



Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Hochschule für Technik und Wirtschaft

Fachbereich Wirtschaftswissenschaften II

Studiengang Angewandte Informatik

Masterarbeit

**Konzeption und Prototyp einer
intelligenten Arbeitssuche nach
persönlichen Fähigkeiten**

vorgelegt von Sergej Meister
Matrikelnummer: s0521159

Erstgutachter: Prof. Dr. Christian Herta

Zweitgutachter: Prof. Dr.-Ing. habil. Dierk Langbein

Berlin, 10. Februar 2016

Vorwort

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation	1
1.2	Ziele der Arbeit	2
1.3	Aufbau der Arbeit	3
2	Grundlagen	4
2.1	Dokumentenorientierte Datenbanken	4
2.2	IntelliJob	7
2.2.1	Technologie Stack	9
2.2.2	Software Architektur	10
2.2.3	Problembeschreibung	11
2.3	Tag Cloud	11
2.4	Informationsbeschaffung	12
2.4.1	Qualität in Information Retrieval System	13
2.4.2	Indexierung von Textdokumenten	14
2.4.3	Boolesche Suche	14
2.4.4	Vektorraummodell	15
2.4.5	Volltextsuche	16
2.4.6	Gewichtete Suche	17
2.4.7	Autovervollständigung	17
2.5	ElasticSearch	17
2.5.1	Sharding	19
3	Systementwurf	21
3.1	OpenNLP oder ElasticSearch	21
3.2	Prozessablauf	21
3.3	Datenstruktur	22

3.4 Schichtenarchitektur	22
4 Entwicklung und Implementierung	23
4.1 Elasticsearch - Konfiguration	23
4.2 Benutzerprofile	23
4.2.1 Modell	24
4.2.2 Controller	24
4.2.3 Services	24
4.2.4 View	25
4.3 Volltextsuche	25
4.3.1 Modell	25
4.3.2 Controller	25
4.3.3 Services	26
4.3.4 View	26
4.4 Gewichtete Suche	26
4.4.1 Modell	27
4.4.2 Controller	27
4.4.3 Services	27
4.4.4 View	27
4.5 Autovervollständigung	28
4.5.1 Modell	28
4.5.2 Controller	28
4.5.3 Services	29
4.5.4 View	29
5 Evaluierung	30
5.1 Volltext-Suchanfragen	31
5.2 Gewichtete Suchanfragen	40
5.3 Vergleich und Fazit	43
6 Problemen und Schwierigkeiten	45
7 Zusammenfassung und Ausblick	46
Literaturverzeichniss	i
Abbildungsverzeichnis	ii

Tabellenverzeichnis	iii
Glossar	iv
Anhang	v
Eigenständigkeitserklärung	xviii

1. Einleitung

1.1 Motivation

Was versteht man unter dem Begriff *Information*? Was sind *Daten*? Die beiden Fragen sind sehr beliebt in der Informatikstudium, besonders im ersten Semester. Doch die Informationen und Daten spielen nicht nur in der Informatik sondern in allen Gebieten eine wichtige Rolle. Täglich treffen die Menschen eine Entscheidung auf der Basis Ihres Wissensdaten, aus denen Informationen extrahiert werden. Dadurch dass Informationen aus den Daten gelesen werden, hängt die Entscheidungsfaktor wesentlich von dem „*Qualität*“ den bereitgestellten Daten ab.

In der heutigen Informationsgesellschaft durch das schnelle Wachstum von digital verfügbaren Daten ist es für die Menschen schwieriger geworden, die entscheidungsrelevante Informationen aus den Datenmengen zu filtern. Deswegen bietet das Internet viele verschiedene Suchportals für unterschiedliche Interessengruppe, die eine gezielte Suche nach bestimmten Datenmengen erlauben. Einer davon sind Job Portals.

Das Job Portal unterstützt ihrer Anwender rund um das Thema „*Arbeitsuche*“ und „*Bewerbungsprozess*“. Ein Arbeitssuchender legt Suchkriterien, wie Berufsbezeichnung und Region, fest und bekommt in weniger Sekunde eine Liste von Arbeitsangeboten, die den Suchkriterien entsprechen. Es ist offensichtlich, dass die Arbeitssuche dank Job Portals viel effizienter und einfacher geworden ist. Trotzdem muss der Arbeitssuchender immer noch viele Informationen selbst raus finden und interpretieren. Zum Beispiel für die Entscheidung, ob die Stelle den gewünschten Anforderungen entspricht, sind die Daten wichtig, wie Berufsname, Firmenname und Qualifikation. Für das Bewerbungsschreiben werden noch weitere Daten benötigt, wie Kontaktperson, Mailadresse, Firmenhomepage und Firmenadresse. Diese Informationen muss der Arbeitssuchender ständig selbst rausfinden. Genau mit dieser Problematik beschäftigt sich das Programm „*Intellijob*“, das diese Daten automatisch aus dem Arbeitsangebot extrahiert.

Das Programm „*Intellijob*“ wurde im Rahmen der Veranstaltung „*Informationssystemen*“ unter Betreuung von Professor Christian Herta entwickelt und kann bereits Berufsbezeichnung, Mailadresse, Firmenhomepage, Kontaktperson und Firmenadresse automatisch auslesen und in Datenbank speichern. Außerdem verfügt das Programm über eine Weboberfläche, die die gespeicherte Daten tabellarisch darstellt.

Der weitere Nachteil von Job-Portals ist die Implementierung von der Suche, die die in der Profile gespeicherte personalisierte Daten nicht berücksichtigt. Denn die typische Suchkriterien eines Job Portals sind einfache Stichworte wie Softwareentwickler, Java, C# u.s.w. Daraus wird es ersichtlich, dass die Suche nach einem konkreten Beruf erfolgt. Dieses Suchverfahren eignet sich gut für Quereinsteiger und für die Menschen, die auf der Suche nach neuer Berufsorientierung sind, aber nicht für die Menschen, die eine Arbeit entsprechend ihren Fähigkeiten und Qualifikationen suchen. Wäre es nicht besser, wenn die Suche nach persönlichen Fähigkeiten und Qualifikationen erfolgte, die von Benutzer selbst bewertet werden? Nehmen wir an, der Mensch bewertet seine Java-Kenntnisse mit 2 und C# mit 5 Sternen. Dann ist es doch logisch, dass alle C# Arbeitsangebote in der Ergebnisliste zuerst angezeigt werden müssen. Vielleicht ist es auf dem ersten Blick nicht ganz ersichtlich, woran die beide Verfahren sich unterscheiden. Die Suche nach einem Beruf als Softwareentwickler kann aber die Java- Arbeitsangebote zuerst anzeigen, was aus der Sicht der Arbeitssuchender mit den besseren C# Kenntnissen nicht ganz zutreffend ist. Der weitere Unterschied liegt im Entscheidungsfaktor. Bei dem ersten Verfahren entscheidet der Mensch, welche Arbeit er haben will. Bei dem zweiten Verfahren entscheidet das Programm, welche Arbeit für den Menschen am besten geeignet ist.

Die Untersuchung von mehreren Job-Portals wie „Xing“, „StepStone“, „JobScout24“, „Monster“ hat bestätigt, dass es tatsächlich kein deutsches Job-Portal gibt, das die Suche nach persönlichen Fähigkeiten und Qualifikationen unterstützt.

1.2 Ziele der Arbeit

Im Rahmen dieser Masterarbeit wird ein Suchverfahren implementiert, dass die Suche nach persönlichen Fähigkeiten und Qualifikationen unterstützt. Die Suche muss mit *Elasticsearch*¹ realisiert und in der Anwendung *IntelliJob* integriert werden. Um die Suchqualität bewerten zu können, müssen mehrere unterschiedliche Suchqueries gebildet und genau untersucht werden. Am Ende der Arbeit wird geprüft, ob das neue Suchverfahren gegenüber der traditionelle Suche nach Stichworte eine bessere Informationsmenge liefert. Das Suchverfahren wird als effektiv bewertet, wenn die erste Arbeitsangebote in der Ergebnisliste den persönlichen Fähigkeiten und Qualifikationen besser entsprechen werden.

Das weitere Thema dieser Arbeit wird die Datenbeschaffung sein. Um nach persönliche Fähigkeiten und Qualifikationen suchen zu können, müssen diese Daten zuerst erfasst werden. Dafür wird das Projekt „IntelliJob“ um eine neue Web-Form erweitert, die die Bewertung von persönlichen Fähigkeiten und Qualifikationen ermöglicht. Die Daten, die dem Arbeitssuchender bereitgestellt werden, müssen sinnvoll gruppiert werden, so dass die Kriterien wie Java und C# unter den Sammelbegriffen Programmieren, Softwareentwickeln, gefunden werden könnten. Der Arbeitssuchender muss in der Lage sein, seine eigene Daten speichern, editieren

¹www.elastic.co.

und löschen zu können. Außerdem darf der Benutzer selbst entscheiden, welches Suchverfahren eingesetzt werden soll.

1.3 Aufbau der Arbeit

Der schriftliche Teil der Arbeit beschreibt das komplette Entwicklungsprozess. Im nächsten Kapitel „*Grundlagen*“ werden die Technologie erläutert, die für die Entwicklung neues Suchverfahrens relevant sind. Dabei wird das Projekt *IntelliJob* und die eingesetzte Technologie kurz vorgestellt. Das Kapitel „*Systementwurf*“ befasst sich mit Softwarearchitektur und beschreibt die Logikverteilung und die Datenkommunikation innerhalb der Anwendung. Das vierte Kapitel „*Entwicklung und Implementierung*“ stellt das Prototyp des neuen Suchverfahrens vor. Dabei wird sowohl das Datenmodel als auch die wichtigsten Funktionen detailliert beschrieben. Außerdem wird in diesem Kapitel auch die API-Schnittstelle dokumentiert. Im Unterkapitel „*Benutzeroberfläche*“ wird die Web-Form zur Erfassung von persönlichen Fähigkeiten und Qualifikationen als Bild dargestellt. Das Kapitel „*Evaluierung*“ analysiert mehrere unterschiedliche Suchanfragen und bewertet sie. Problemen und Schwierigkeiten, die während der Implementierung aufgetreten sind, sowie die Art und Weise, wie diese Probleme gelöst oder eventuell nicht gelöst wurden, werden in dem 6. Kapitel erfasst. Zum Schluss wird die Arbeit zusammengefasst und ein Resümee gezogen. Ein Ausblick zeigt auf, welche Erweiterungen für die Anwendung in Zukunft noch sinnvoll sein könnten.

2. Grundlagen

In diesem Kapitel wird auf die allgemeinen und informatischen Grundlagen eingegangen, welche zum Verständnis dieser Masterarbeit relevant sind. Die Anwendung *IntelliJob* persistiert bereits die Daten in eine dokumentenorientierte Datenbank *MongoDB*¹ und es wird noch eine weitere dokumentenorientierte Technologie *Elasticsearch* eingesetzt. Deswegen ist es wichtig mindestens die Grundlagen sowie die Vor- und Nachteile von dokumentenorientierten Datenbanken im Blick zu behalten. Die in der Arbeit verwendete *Elasticsearch-Feature's* werden in einem extra Kapitel ausführlich beschrieben. Da die Suchfunktionalität in der bereits vorhandene Anwendung integriert werden muss, ist es für die Bewertung der Arbeit notwendig den IST-Zustand zu dokumentieren.

2.1 Dokumentenorientierte Datenbanken

Bei dokumentenorientierten Datenbanken werden die Daten in Form von Dokumenten gespeichert. Die typische Dokument-Formate sind XML, YAML und JSON. Die einzelnen Datenfelder werden als Key-Value Stores zusammengefasst. Jedes Dokument ist vollkommen frei bezüglich seine Struktur und das ist ein wesentlicher Unterschied zu relationalen Datenbanken, wo die Datenstruktur vordefiniert ist [Dorschel 2015]. Die Abbildung 2.1 zeigt eine typische normalisierte Datenstruktur in den relationalen Datenbanken. Einige Datensätze in der Tabelle *skills* beinhalten kein "*description*". Die Spalte muss in dem Fall einfach leer bleiben. Es besteht keine Möglichkeit nur für bestimmte Datensätze auf die Spalte "*description*" zu verzichten.

category_id	name	skill_id	category_id	name	description
1	Languages	1	1	German	
2	Knowledges	2	1	English	
3	Personal strengths	3	2	Java	programming language java
		4	2	PHP	programming language php

Abbildung 2.1: Relationale Tabellen *skill-categories* und *skills*

In dokumentenorientierten Datenbanken ist es aber durchaus möglich, dass einige Dokumente innerhalb einer *Collection* kein *description* Feld haben. Das sieht man in der *Collection skills* auf der nächsten Seite.

¹www.mongodb.org.

Mongo Collection *skill-categories*

```
{
  "_id": "569ab65644ae6344028b6cb7",
  "name": "Languages"
},
{
  "_id": "569ab65644ae6344028b6cb7",
  "name": "Knowledges"
},
{
  "_id": "569ab66d44ae6344028b6cb9",
  "name": "Personal strengths"
}
```

Mongo Collection *skills*

```
{
  "_id": "569ab85144ae6344028b6cbf",
  "category": {
    "_id": "569ab61a44ae6344028b6cb5",
    "name": "Languages"
  },
  "skills": [
    {
      "_id": "569ab8dd44ae6344028b6cc0",
      "name": "deutsch"
    },
    {
      "_id": "569ab8dd44ae6344028b6cc1",
      "name": "english"
    }
  ]
},
{
  "_id": "569ab9d644ae6344028b6cc9",
  "category": {
    "_id": "569ab65644ae6344028b6cb7",
    "name": "Knowledges"
  },
  "skills": [
    {
      "_id": "569ab9d644ae6344028b6cca",
      "name": "java",
      "description": "programming language java"
    }
  ]
}
```

Wenn man die beiden *Collection* genau anschaut, fehlt es sofort auf, dass die *Category-Daten* nicht redundant sind. In den relationalen Datenbanken werden die Datenattribute durch die Anwendung von verschiedenen Normalisierungsregeln so lange zerlegt, bis keine vermeidbaren Redundanzen mehr vorhanden sind. Die dokumentenorientierte Datenbanken verfolgen den Prinzip der *Aggregation*. Der Begriff kommt eigentlich aus dem Domain-Driven Design und bedeutet, dass jedes Dokument eine eigenständige Einheit ist, das von keinen anderen beziehungsweise von möglichst wenigen Dokumenten abhängig ist. Dadurch können die Daten immer vollständig automar gespeichert oder gelesen werden.

Diese hohe Datenflexibilität ist aber nicht umsonst. Zum Ersten ist es oft aufwendig die Daten zu ändern, die bereits in anderen Dokumenten gespeichert sind. Ein Beispiel wäre dafür das Dokument mit dem Name „*Languages*“ in der *Collection* „*skill-categories*“ mit einem neuen Attribute zu erweitern oder komplett zu löschen. In dem Fall muss diese „*Category*“ auch in allen anderen Dokumenten gefunden und entsprechend geändert werden. Zum Zweiten können dokumentenorientierte Datenbanken aufgrund ihrer Schemafreiheit keine einfache Datenvalidierung der Attributen und Datentypen durchführen. Das muss stattdessen im Programmcode behandelt werden, was zu höheren Softwarekosten führt und fehleranfällig ist [Jan 2013].

Wenn man das ganze kurz wörtlich zusammenfasst: Schemafreiheit, keine Datenvalidierung, mehr Programmcode, Fehleranfällig und aufwändige Datenänderung auf der Datenbankebene, fragt man sich, wieso werden aber die dokumentenorientierte Datenbanken, die seit 2008 unter dem Titel „*NoSQL*“² bekannt sind, immer beliebter? Eine einfache Antwort lautet, dass die NoSQL-Datenbanken oft schneller als relationale Datenbanken sind. Statt mehrere Tabellen mit JOIN-Operation abfragen zu müssen, führt ein NoSQL-Datenbank nur eine einzige Abfrage aus, um ein kompletter Datensatz zu erhalten. Die Datenkonsistenz beim Lesen und Schreiben in mehrere Tabellen wird in klassischen relationalen Datenbanken durch Transaktion-Mechanismen realisiert. Jede Transaktion ist isoliert von einander und wird ganz oder gar nicht ausgeführt. Wenn eine Transaktion in eine Tabelle schreibt, kann keine andere Transaktion auf die Tabelle zugreifen, bis der Schreibvorgang komplett abgeschlossen oder unterbrochen wird. Das führt dazu, dass die Lese- und Schreiboperation von parallel ausgeführten Transaktionen einander blockieren und warten müssen, bis alle nötige Ressource freigegeben werden. Dadurch dass NoSQL Datenbanken keine Abfrage blockieren, werden die auch schneller abgearbeitet. Es besteht zwar die Möglichkeit, dass inkonsistente Daten gelesen werden, aber aufgrund, dass die Daten oft nur aus einer einzigen *Collection* gelesen werden, ist die Fehlerwahrscheinlichkeit enorm klein im Vergleich zu viel steigender Performance[Dorschel 2015]. Die einzige Schutzmaßnahme, die von meisten NoSQL-Datenbanken unterstützt werden, ist die Datenversionierung. Bei jedem Speicherzugriff wird die Version des Dokumentes geprüft, ist die Version identisch dem bereits gespeicherten Dokument, dann wird ein *Exception* geworfen. Denn es ist ein Zeichen dafür, dass Dokument bereits von einem anderen *Request*

²<https://de.wikipedia.org/wiki/NoSQL>.

geändert wurde. Außerdem wird auch die Datenversionierung für die Synchronisierung zwischen mehrere Rechnerknoten innerhalb eines *Clusters* verwendet. Das *Sharding Concepts* in dokumentenorientierten Datenbanken wird im Kapitel über *Elasticsearch* detailliert beschrieben.

Fazit: Die dokumentenorientierte Datenbanken eignen sich besonders gut für die Anwendungen, wo hohe Geschwindigkeit bei Datenzugriffe und Datenverarbeitung im Vordergrund steht und die Datenredundanz nicht hoch priorisiert wird. Die meisten Datenbankzugriffe in der Anwendung *IntelliJob* sind Lesezugriffe. Es gibt wenige Möglichkeiten die Daten durch Benutzeraktion zu ändern. Die Änderungen beeinflussen nur eine einzige *Collection* und oft sogar nur ein einziges Dokument. Deswegen ist die Entscheidung über den Einsatz von dokumentenorientierten Datenbanken richtig.

2.2 IntelliJob

Die Anwendung „*IntelliJob*“ wurde bereits mehrmals erwähnt und muss endlich vorgestellt werden. Es gab mal eine Idee die bewerbungsrelevante Daten aus Arbeitsangeboten automatisiert zu extrahieren. Diese Idee wurde mit dem Programm „*IntelliJob*“ realisiert. Da die Anwendung im Rahmen der Veranstaltung „*Informationssystemen*“ eigenständig entwickelt und dokumentiert wurde, werden einige Inhalte aus der Arbeit in diesem Kapitel zitiert.

Bevor die Daten aus dem Arbeitsangebot gelesen werden könnten, muss dieses Arbeitsangebot erstmals ermittelt werden. Dafür werden mehrere Job-Agenten auf Job-Portals *Stepstone*³ und *Monster*⁴ angelegt, die die Arbeitsangebote per Email täglich senden. Das Email beinhaltet mehrere HTML-Links mit unterschiedlichen Arbeitsangeboten. Die Anwendung „*IntelliJob*“ holt diese Emails aus dem Postfach ab, ruft die Links auf und speichert das geladene Inhalt in Datenbank ab.

Im nächsten Schritt werden die bewerbungsrelevante Daten, *Berufsbezeichnung*, *Firmenadresse*, *Firmenhomepage*, *Mailadresse* und *Kontaktperson*, aus dem Arbeitsangebot extrahiert.

Berufsbezeichnung wird einfach aus dem HTML-Link ausgelesen, das als *Value* immer eine Berufsbezeichnung enthält.

Mailadresse und **Firmenhomepage** werden mit Hilfe von Regulären Ausdrücken extrahiert.

Mail-Pattern: `[a-zA-Z0-9_+.]+@[a-zA-Z0-9-]+.[a-zA-Z0-9-]+`

WWW-Pattern: `www.[\w\d.:#@%/$()~_?+&]*`

HTTP-Pattern: `http.[\w\d.:#@%/$()~_?+&]*`

³www.stepstone.de.

⁴www.monster.de.

HTTPS-Pattern: `https.[\w\d.:#@%/\;$()~_?+&]*`

Leider können Reguläre Ausdrücke nur Daten mit einer eindeutigen Struktur ermitteln. **Kontaktperson** und **Firmenadresse** können aber sehr unterschiedlich geschrieben werden:

- Alle der Kosmonauten 26a D-32451 Bad Kreuznach
- Straße des 17. Juni 17a 13456 Berlin (Lieblings Beispiel)
- Herr Richard-Alexander Wagner
- Johann Wolfgang von Goethe
- Alexander Sergejewitsch Puschkin

Aus den oben aufgelisteten Beispielen wird es ersichtlich, dass es kaum möglich ist, allein mit Regulären Ausdrücke, diese Daten zu extrahieren. Deswegen werden diese Daten in „IntelliJob“ mit Hilfe von „Apache OpenNLP Framework“⁵ gefunden und zusammen mit allen anderen bewerbungsrelevanten Daten in Datenbank gespeichert.

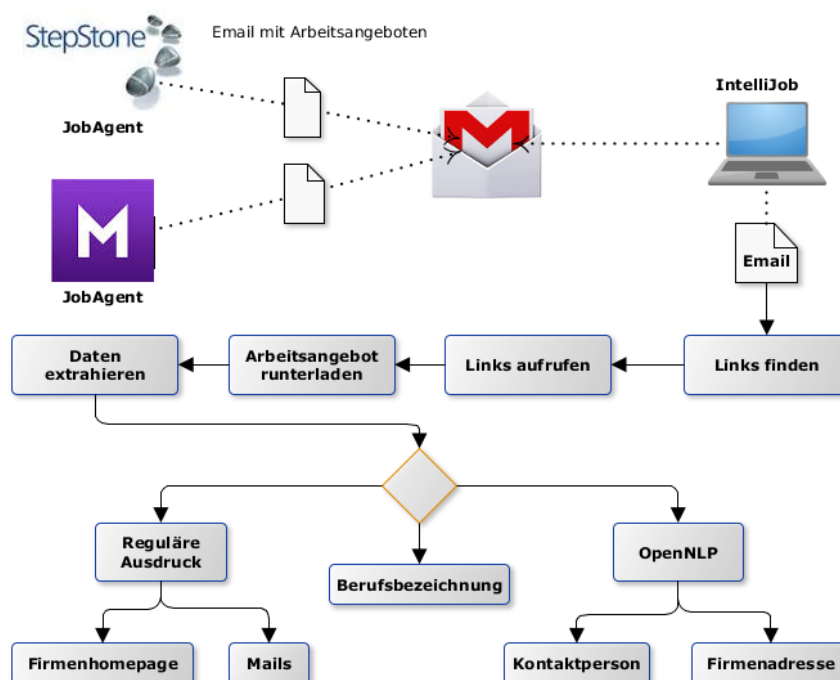


Abbildung 2.2: Prozessablauf in IntelliJob

⁵<https://opennlp.apache.org>.

2.2.1 Technologie Stack

Apache OpenNLP ist ein Open Source Produkt und wird als Maven-Dependency in der Version 1.5.3 in das Projekt „*IntelliJob*“ eingebunden. Das Framework setzt die Technologie aus dem Gebiet *Natural Language Processing* ein, um bestimmte Bestandteile eines Textes zu erkennen und zu klassifizieren. Die Muster-Erkennung basiert auf künstlich erzeugten Trainingsdaten, die durch ein Lernprozess in einem Model gespeichert werden. Das Framework wird ständig weiter entwickelt und verfügt über eine gute Dokumentation.⁶

Die Anwendungslogik von „*IntelliJob*“ wurde in der Programmiersprache Java geschrieben und verwendet sehr viele Komponenten aus dem Spring Framework⁷. Spring Framework ist ebenfalls ein Open Source Produkt, welches performante und moderne Java-Enterprise Einsätze realisiert. Mit dem Einsatz von Spring *Dependency Injection* und *Aspekt orientierte Programmierung* wird das Programmcode redundant verteilt und kann auch leicht wiederverwendet werden. Alle erzeugte Objekte werden anhand der *Spring Data*⁸ in eine dokumentenorientierte Datenbank MongoDB persistiert.

Die weitere Besonderheit von „*IntelliJob*“ ist, dass es eine Standalone Web-Anwendung ist, die lokal auf dem Client einen embedded Web Container *Tomcat*⁹ startet. Der große Vorteil von diesem Einsatz liegt darin, dass der Webserver ein Teil der Anwendung ist, und die Anwendung nicht mehr in den Webserver deployed werden muss. Das Programm wird als JAR-Datei ausgeliefert und kann direkt in der Java Virtual Machine ausgeführt werden. Realisiert wird das ganze mit *Spring Boot*¹⁰ Technologie, die dank Autokonfiguration sowohl alle nötige Resource zusammenpackt, als auch alle Rest-Services zur Kommunikation mit Frontend initialisiert.

Das Frontend wird mit *AngularJS*¹¹ umgesetzt und verfügt bereits über 6 verschiedene Views. Das Framework unterstützt die Entwickler bei der Implementierung von Single Page Anwendungen.

Home	Startseite um Mail-Account auszuwählen und Zugangsdaten zu übergeben.
Emails	zeigt alle gefundene Emails von Job Portals
Job Links	zeigt alle in Mails gefundene Links zu Arbeitsangeboten
Jobs	zeigt alle runter geladene Arbeitsangebote
Jobs Details	zeigt alle extrahierte Daten eines Arbeitsangebotes
Audit	dient zur Datenauswertung und zeigt sowohl die aktuellste als auch die alte Ergebnisse des Datenextraktions.

Während der Masterarbeit wird *AngularJS* von der Version 1.2.16 zu 1.4.7 aktualisiert.

⁶<https://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html>.

⁷www.spring.io.

⁸<http://projects.spring.io/spring-data>.

⁹<http://tomcat.apache.org>.

¹⁰<http://projects.spring.io/spring-boot>.

¹¹<https://angularjs.org>.

2.2.2 Software Architektur

„IntelliJob“ ist quelloffen und kann auf Versionsverwaltungsplattform, GitHub¹², heruntergeladen werden. Die Anwendungslogik basiert auf Schichtenarchitektur, die sequenziell von oben nach unten abgearbeitet wird.

Präsentationsschicht repräsentiert die Daten in HTML-Form und regelt die Interaktionen zwischen Benutzer und Software. Alle Benutzeranfragen werden asynchron an Serviceschicht weitergeleitet. Die Daten zwischen beiden Schichten werden in JSON-Format transportiert.

Serviceschicht nimmt die Client-Anfragen entgegen, validiert Benutzerdaten und wandelt sie in Domain-Objekte um. Danach wird die Abfrage der Controlschicht übergeben. Letztendlich wird ein *Response* generiert und zurückgesendet.

Controlschicht vereint alle Methoden aus Datenzugriffsschicht und führt alle nötige Operationen aus, um die Benutzeranfrage zu bearbeiten.

Datenzugriffsschicht kapselt die Zugriffe auf persistente Daten. Die Datenaustausch erfolgt über Domain-Objekte.

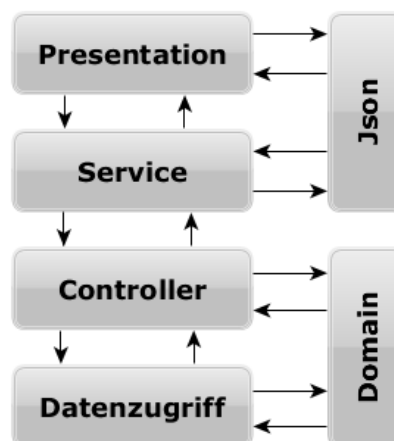


Abbildung 2.3: Schichtenarchitektur

Dadurch dass die Anwendungslogik in mehrere Schichten verteilt wird, wird die Komplexität der Anwendung wesentlich reduziert, was sowohl für das Verständnis als auch für die Wartung von Software eine große Vorteil ist. Außerdem können die Programmteile dank seiner Abstraktion gut getestet und auch leicht wiederverwendet werden.

Um die Methoden zum Datenextraktion auch in anderen Projekten zu verwenden, werden sie in ein separates Projekt, namens *Civis-Tools*¹³ ausgelagert.

¹²<https://github.com/SergejMeister/intellijob>.

¹³<https://github.com/SergejMeister/civis-tools>.

entspricht und verlässt die Seite. Da die Seite jedoch aufgerufen wurde, wird der Schlagwort der Tag Cloud höher gewichtet[OnPage 2016].

Die Tag Cloud wird meistens nach der folgenden Formel berechnet:

$$S_i = [(f_{max} - f_{min}) * \frac{t_i - t_{min}}{t_{max} - t_{min}} + f_{min}]$$

- S_i - anzuzeigende Schriftgröße
- f_{max} - maximale Schriftgröße
- f_{min} - minimale Schriftgröße
- t_i - Häufigkeit des betreffenden Schlagwortes
- t_{min} - Häufigkeit, ab der ein Schlagwort angezeigt werden soll
- t_{max} - Häufigkeit des häufigsten Schlagwortes¹⁴

2.4 Informationsbeschaffung

Im ersten Kapitel wurde schon erwähnt, dass die Informationen aus Daten extrahiert werden. Zur Informationsbeschaffung (eng. Information Retrieval) gehören neben Extraktionsprozess noch weitere Prozesse. Zuerst müssen die Daten erfasst werden und in einer Datenbank oder einem Daten-Management-System in einer bestimmten Syntax abgelegt werden. Dabei legt der Verfasser einen Teil seines Wissen in ein Dokument nieder, das in Text, Bild oder eine andere Form gespeichert wird. Im nächsten Schritt müssen die Daten, die bereits existieren, wiedergefunden und in geeignete Repräsentationsform gebracht werden. Letztendlich sollen nur die Informationen extrahiert werden, die möglichst gut bei einer Problemlösung helfen können. Der Autor Andreas Heinrich in seinem Lehrtext *„Information Retrieval Grundlagen, Modelle und Anwendungen“* bezeichnet diesen Prozess als Wissenstransfers und weist deutlich darauf hin: *„... Information Retrieval nicht nur als den Prozess der Wissensextraktion aus einer Dokumentensammlung sondern als Prozess des Wissenstransfers zu verstehen“*[Henrich 2008]. Die Abbildung 2.7 stellt das Prozess vereinfacht dar.

Das System, das sich mit der Informationsbeschaffung befasst, wird in der Informatik als IR-System bezeichnet. Die Aufgabe des IR-Systems besteht hauptsächlich darin, die hinter der Anfrage stehende Informationsbedürfnisse zu befriedigen. Oft muss ein IR-System neben der Anfrage noch weitere Datenquellen, wie Feedback oder vorhandene Information über Nachfragender, zu berücksichtigen, um ein besseres Informationsergebnis zu erzielen. Im Bezug auf die Anwendung *„IntelliJob“* kann Informationsergebnis wesentlich verbessert werden, indem der Nachfragender die Suchanfrage selber priorisiert.

¹⁴<https://de.wikipedia.org/wiki/Schlagwortwolke>.



Abbildung 2.5: Prozess des Wissenstransfers vom Autor eines Dokumentes bis zum Informations-Nachfragenden

2.4.1 Qualität in Information Retrieval System

Um ein gutes IR-System von einem schlechten unterscheiden zu können, muss die Qualität der Ergebnisse gemessen werden. Die Qualität wird ihrerseits durch **Relevanz**, **Recall** und **Precision** bestimmt.

Relevanz kann binär (*relevant/nicht relevant*) oder mit Relevanzeinstufung (von 1 bis 10) definiert werden. In der Praxis hat sich allerdings die binäre Relevanz durchgesetzt. Die Entscheidung, ob ein Dokument relevant oder nicht relevant ist, ist sehr subjektiv und erfordert Vorwissen sowohl über das untersuchte Dokument als auch über andere Dokumenten. Deswegen wird es sehr schwierig alle Dokumente einzustufen [Henrich 2008].

Recall beschreibt der Anteil der relevanten Dokumenten in der Ergebnismenge im Verhältnis zum allen relevanten Dokumenten und wird wie folgt berechnet:

$$Recall = \frac{a}{S}$$

- a - Anzahl der relevanten Dokumente im Ergebnis

- S - Gesamtzahl der relevanten Dokumente

Precision zeigt die Genauigkeit der Ergebnismenge.

$$Precision = \frac{a}{D}$$

- a - Anzahl der relevanten Dokumente im Ergebnis
- D - Gesamtzahl der Dokumente im Ergebnis

Recall und **Precision** dürfen aber nicht einzeln betrachtet werden. Denn nur der Vergleich von beiden Werten weist deutlich darauf hin, ob ein IR-System besser oder schlechter gegenüber eines anderen Systems ist.

2.4.2 Indexierung von Textdokumenten

Die vorliegende Arbeit beschäftigt sich mit Arbeitsangeboten, die in Textform vorliegen. Diese Textdokumente müssen automatisiert erfasst werden. Dieses Prozess wird in Information Retrieval als *Indexierung* bezeichnet. Bei der Indexierung wird ein Dokument analysiert und alle Wörter außer *Stopwörter* werden extrahiert. Die *Stopwörter* sind besondere Wörter die sehr häufig vorkommen. In der deutschen Sprache sind Artikeln ein typisches Beispiel für die *Stopwörter*. Die extrahierte Worte werden in Grundform (z.B. *schreiben* - *schreib*) gebracht und als Terme abgespeichert. Durch Grundformredaktion wird es verhindert, dass die gleiche aber unterschiedlich geschriebene Worte mehrmals gezählt werden. Denn während der Indizierung wird sowohl das Vorkommen eines Wortes in einem Dokument (term frequency, TF) als auch in allen anderen Dokumenten (inverse document frequency, IDF) gezählt.

$$IDF_i = \log \frac{N}{n_i}$$

- i - der untersuchte Term
- N - Anzahl aller Dokumenten
- n_i - Anzahl der Dokumente mit Term

Der Logarithmus sorgt dafür, dass ein oft vorkommender Term nicht extrem hoch gewichtet wird. Das Endergebnis der automatischen Indexierung ist ein Gewicht für jeden Term in jedem Dokument [Mandl 2001]. Nachdem das Dokument erfolgreich indexiert wurde, kann es auch durchgesucht werden.

2.4.3 Boolesche Suche

Die Boolesche Suche oder das Boolesche Retrieval Modell basiert auf die elementare Mengenoperationen, wie Schnittmenge (AND-Verknüpfung), Vereinigungsmenge (OR-Verknüpfung) und Komplementärmenge (NOT-Verknüpfung). Das Modell

geht davon aus, dass das Ergebnis einer Suchanfrage als exakte Menge von relevanten Dokumenten bestimmt werden kann. Dabei werden nur Dokumente ausgegeben, die den in der Suchanfrage formulierten booleschen Ausdruck erfüllen.

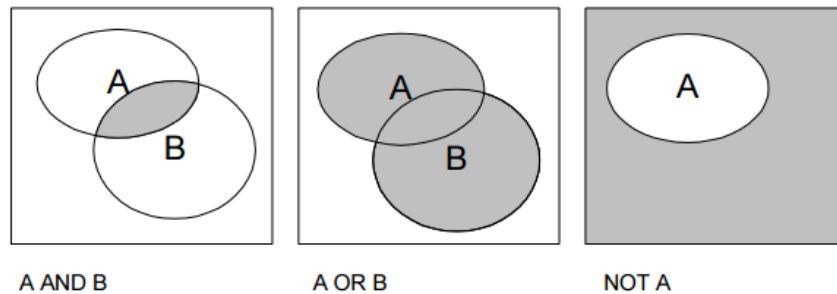


Abbildung 2.6: Die elementaren booleschen Operationen

Der Nachteil des booleschen Modells besteht darin, dass die Gewichtung von Termen nicht berücksichtigt wird und die Ergebnismenge deswegen ungeordnet ausgeliefert wird. Außerdem werden viele Dokumente nicht gefunden, weil die gesuchte Elemente nicht in den exakt gleichen Form in Dokumenten enthalten sind. Trotz seiner Ungenauigkeit wird das boolesche Model oft eingesetzt vor allem um mehrere Suchanfragen mit einander zu verknüpfen.

2.4.4 Vektorraummodell

Zum ersten Mal wurde das Vektorraummodell von Gerard Salton, Andrew Wong und Chungshu Yang in einem Zeitschriftartikel unter dem Thema „*A Vector Space Model for Automatic Indexing*“ im Jahr 1975 publiziert und ausführlich beschrieben[Salton; Wong; Yang 1975]. Die Grundidee dieses Modells besteht darin, dass sowohl die Anfrage als auch alle Dokumente in einem vieldimensionalen Vektorraum dargestellt werden. Dabei bildet jeder einzelne Term eine Achse beziehungsweise eine Dimension in diesem Vektorraum und die Koordinaten werden durch die Gewichtung von Termen repräsentiert.

Die Relevanz eines Dokumentes im Vektorraummodell wird oft durch Berechnung des Cosinus des Winkels zwischen Anfrage und Dokument bestimmt. Das führt dazu, dass nicht mehr nach exakte Übereinstimmungen zwischen Suchanfrage und Dokumenten sondern nach ihre Ähnlichkeit gesucht wird und die Ergebnismenge auch sortiert ausgegeben wird.

Obwohl das Vektorraummodell die wesentliche Nachteile des booleschen Modells aufhebt, kann das boolesche Model unter bestimmte Bedingungen eine bessere Ergebnismenge als Vektorraummodell liefern. Damit die Ähnlichkeitsmaß zwischen Anfrage und Dokument mit Vektorraummodell sinnvoll berechnet werden könnte, muss die Anfrage aus mehrere Begriffe bestehen. Wenn die Suchanfrage aus wenigen Begriffen besteht, kann das boolesche Model, das diese Begriffe einfach mit *AND* verknüpft, oft zu besseren Ergebnissen führen.

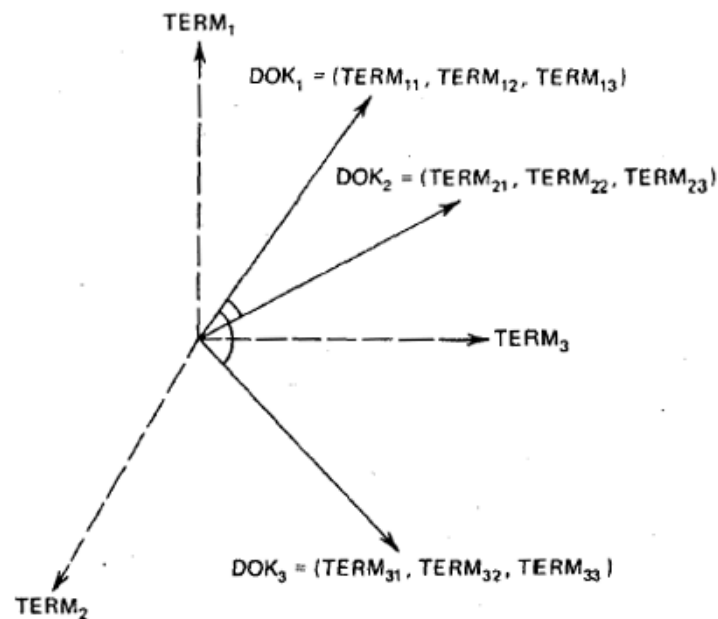


Abbildung 2.7: Vektorraumpräsentation eines Dokumentraums (Salton u. McGill 1987, 129)

2.4.5 Volltextsuche

Bei der Volltextsuche handelt es sich darum, eine benutzerdefinierte strukturierte Suchanfrage in einem unstrukturierten Textdokument zu finden. Natürlich kann man dafür das boolesche oder Vektorraummodell einsetzen, um die erste Ergebnisse zu bekommen. Allerdings bieten die meisten IR-System noch einige Methode für die Optimierung von Suchanfragen, die die Ergebnismenge wesentlich beeinflussen können. Der nachfolgende Abschnitt beschreibt kurz die Optimierungsmethode, die für den praktischen Teil nützlich sein könnten.

Phrasensuche wird für die Suche nach zusammenhängenden Begriffen eingesetzt. Abhängig von Einsatzszenarien kann die Phrasensuche entweder sehr gute oder sehr schlechte Ergebnismenge liefern. Wenn man nach einen Title oder eine exakte Phrase in einem Dokument sucht, dann bekommt man gute Ergebnisse. Allerdings werden viele potenziell relevante Dokumente ausgeschlossen, weil die gesuchte Phrase in diesen Dokumenten nicht in der gleichen Wortfolge vorliegt. Sucht man beispielsweise nach „*relationale Datenbanken*“, werden Dokumente mit „*relationale Datenbank*“ nicht gefunden.

Fuzzy-Suche sorgt für Fehlertoleranz und erlaubt eine vordefinierte Abweichung zwischen den gesuchten und in Dokument vorhandenen Term. Das bekannteste Algorithmus für die Fuzzy-Suche ist Levenshtein-Distanz¹⁵. Das Algorithmus berechnet die Anzahl von Änderungen, die nötig sind, um ein Wort in ein anderes umzuwandeln. Je wenig Änderungen notwendig sind, desto ähnlicher sind die Worte.

¹⁵<https://de.wikipedia.org/wiki/Levenshtein-Distanz>.

Filter werden verwendet, um die Suchergebnisse nach bestimmte Kriterien einzuschränken. Der meistverbreitete Filter ist Seiten-*Pagination*. Die andere Einsatzmöglichkeiten sind Filter für Zeit- oder Dokumententyp.

2.4.6 Gewichtete Suche

Im Kapitel „*Indexierung von Textdokumenten*“ wurde bereits erwähnt, dass jedes Term von IR-System durch Indexierung gewichtet wird. Allerdings nicht nur Terme sondern auch die Suchanfragen(eng. search query) werden von IR-System gewichtet. Wenn die Suchanfrage aus mehrere einzelnen Suchanfragen besteht, dann wird die Gewichtung für alle einzelnen Anfragen ebenfalls berechnet.

Während die Volltextsuche die bessere Ergebnismenge durch Manipulation von Termen erreichen will, mischt sich die gewichtete Suche direkt in der Berechnung der Gewichtung ein. Dieses Verfahren ist als *Boosting* bekannt und wird meistens zur Anzeige von beliebtesten oder populärsten Dokumenten verwendet.

Es gibt zwei unterschiedliche *Boosting*-Verfahren: **Field-Boosting** und **Term-Boosting**. **Field-Boosting** wird eingesetzt, wenn die Suche mehrere Felder anfragt und ein bestimmtes Feld höher als anderen priorisiert werden soll. Beim **Term-Boosting** wird die Suchanfrage durch Vorkommen eines bestimmten Terms höher gewichtet.

2.4.7 Autovervollständigung

Autovervollständigung ist heutzutage fast in jeder Anwendung vorhanden. Die Möglichkeit, durch die Eingabe von wenigen Zeichen auf vermutlich relevanter Suchbegriff zu kommen, hat sich in der Realität als sehr praktisch erwiesen. Grundsätzlich basiert dieses Verfahren auf die Ergänzung von eingegebenen Anfangsbuchstaben. Trotzdem gibt es immer mehr Anwendungen, wo die eingegebene Buchstaben nicht als Anfangsbuchstaben sondern als einen Suchbegriff interpretiert werden und vermutlich relevante Dokumente sofort zum Auswahl angezeigt werden.

Mit Hilfe von Autovervollständigung wird das Erinnerungsproblem zu einem Wiedererkennungsproblem. Denn es reicht wenige Informationen, um ein gewünschter Suchbegriff zu treffen. Außerdem können auch Tippfehler bei der Autovervollständigung berücksichtigt und vermieden werden.

2.5 ElasticSearch

ElasticSearch ist ein Open-Source Suchserver, welcher auf Apache Lucene Bibliothek basiert. Das Projekt wird in der Programmiersprache Java geschrieben und ermöglicht die Kommunikation über REST-Schnittstelle. Die Daten werden in JSON-Format übertragen und in einem invertierten Index abgespeichert. Ein Index entspricht einem Datenbank im Datenbankmanagement System. Jeder Datensatz

wird durch Dokumenten und Spalten durch Felder repräsentiert. Elasticsearch verfolgt eine dokumentenorientierte Architektur und ist deswegen in der Lage auch komplexe Objekte in einem Feld zu speichern.

Relationale Datenbank	Elasticsearch
Datenbank	Index
Tabelle	Dokumenttyp
Datensatz	Dokument
Spalte	Feld

Tabelle 2.1: Begriffsvergleich zwischen Relationale Datenbanken und Elasticsearch

Elasticsearch unterstützt alle im Kapitel 2.4 vorgestellten Methoden der Informationsbeschaffung und erweitert sie mit neuen Funktionalität. Zum Beispiel ist es möglich die Dokumente in Echtzeit zu indexieren. Die Indexierung wird dabei im Arbeitsspeicher gehalten und erst später auf die Festplatte geschrieben. Besonders mächtig ist Elasticsearch bei der Berechnung von Gewichtung(eng. score). Denn es werden mehrere Funktionen zur Manipulation von Score bereitgestellt. Es ist sogar erlaubt eigene Funktion zur Berechnung von Score zu schreiben. Standardmäßig wird aber Score nach Lucene-Formel berechnet¹⁶.

$$score(q, d) = queryNorm(q) * coord(q, d) * \sum_{t(in)q} (tf * idf^2 * boost(t) * norm(t, d))$$

- $score(q, d)$ - Relevanz eines Dokumentes (d) für die Suchanfrage (q)
- tf - term frequency (Kapitel 2.4.2) - $tf = \sqrt{termFreq}$
- idf - inverse document frequency (Kapitel 2.4.2) - $idf = 1 + \log(\frac{numDocs}{docFreq+1})$
- $queryNorm(q)$ - Funktion zur Berechnung der Normalisierung von *Search-Query* - $queryNorm(q) = 1/\sqrt{sumOfSquaredWeights}$, $sumOfSquaredWeights$ summiert IDF von allen Termen in der Suchanfrage
- $coord(q, d)$ - wird verwendet, um Dokumente mit mehrere vorhandenen Suchtermen höher zu gewichten $coord(q, d) = \frac{t(in)q}{t(in)d}$
- $boost(t)$ - erlaubt *boosting* von *Search-Query* (Kapitel 2.4.6). Standardmäßig ist die Wert gleich 1.
- $norm(t, d)$ - Funktion zur Berechnung von Feldnormalisierung
 $norm(t, d) = 1/\sqrt{numTerms}$, $numTerms$ steht für die Anzahl von Termen im gesuchten Feld

Die wichtige Funktion, die in diesem Zusammenhang auch erwähnt werden soll, ist die *Decay*-Funktion. *Decay*-Funktion berechnet die Relevanz eines Dokumentes abhängig von dem benutzerdefinierten Distanz zwischen einem numerischen

¹⁶www.elastic.co/guide/en/elasticsearch/guide/current/practical-scoring-function.html.

Feldwert und dem Original-Wert. Das numerische Feld kann eine Zahl, ein Datum oder Geo-Location sein. Die Distanz wird *linear*, *exponential* oder standardmäßig nach *Gaus*-Algorithmus berechnet.

$$S(doc) = \exp\left(-\frac{\max(0, |fieldValue - origin| - offset)^2}{2\sigma^2}\right)$$

- *fieldValue* - Feldwert
- *origin* - Startwert, der bei der Distanzberechnung berücksichtigt wird. Beim Datum kann es zum Beispiel das heutiges Datum sein.
- *offset* - Decau-Funktion wird berechnet, wenn die Distanz größer als *offset*. Standardwert ist 0.
- σ^2 -sorgt dafür, dass die berechnete Distanz in benutzerdefinierten Intervall liegt. $\sigma^2 = -\frac{scale^2}{2 \cdot \log(decay)}$, *decay* definiert den Abstand zwischen den Dokumenten. Standardwert ist 0.5. *scale* ist die benutzerdefinierte Distanz.

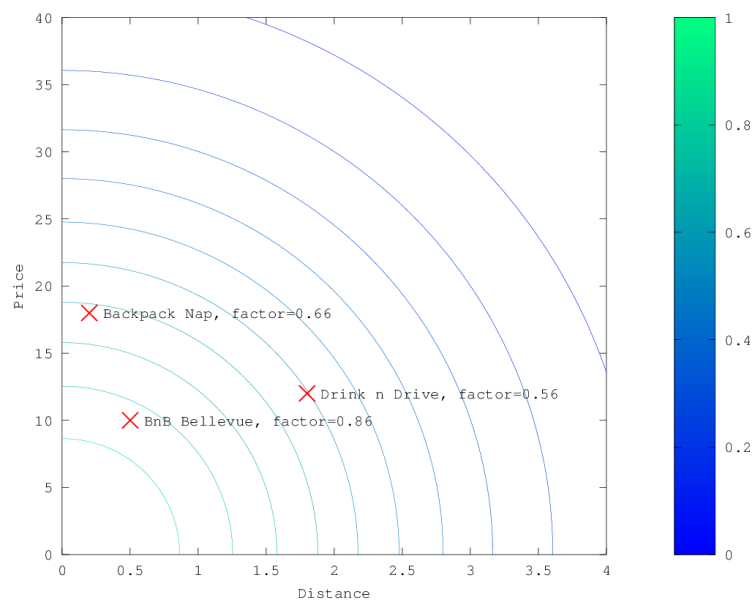


Abbildung 2.8: Visualisierung von *Decay*-Funktion mit *Gaus*

2.5.1 Sharding

Das Wort *Elastic* in dem Projektname bezieht sich auf horizontale Skalierbarkeit, welche ermöglicht, die Daten auf mehrere Server zu verteilen. Alle Daten werden in einem Cluster vorgehalten. Ein Cluster besteht aus einem oder mehreren Knoten (eng. Node), die auf unterschiedlichen Server laufen können. Die Synchronisierung zwischen den Knoten wird über Cluster gesteuert und kann relativ einfach und flexibel von Benutzer konfiguriert werden. Wenn ein Server besonders gut aufgerüstet ist, ist es sinnvoll ein *Master* und *Slave* Knoten festzulegen.

Um Ausfallsicherheit zu gewährleisten, werden die Dokumente der einzelnen Indizes in mehrere *Shards* repliziert und auf den Knoten verteilt. Der Shard, welcher das Dokument zuerst speichert, wird *primary shard* genannt. Alle weitere *shards* sind *replica shards* und behalten nur die Kopie von diesem Dokument. Wenn ein Node ausfällt, kann das Dokument aus dem *replica shard* eines anderen Nodes gelesen werden[Rafal Kuc 2013]. Die Abbildung 2.10 stellt ein Cluster mit zwei Knoten, 4 *primary shard* und 1. *replica shard* dar.

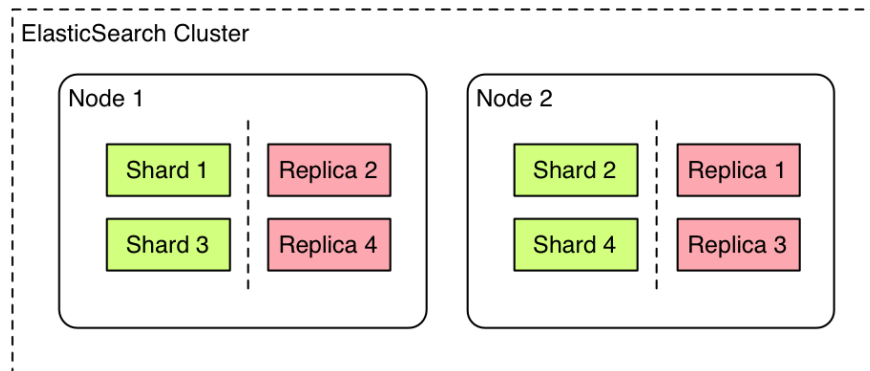


Abbildung 2.9: Beispiel - Ein Cluster mit zwei Knoten, 4 *shards* und 1 *replica shard*[Terrier 2013]

3. Systementwurf

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

3.1 OpenNLP oder ElasticSearch

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

3.2 Prozessablauf

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdspra-

chige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

3.3 Datenstruktur

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

3.4 Schichtenarchitektur

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4. Entwicklung und Implementierung

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.1 Elasticsearch - Konfiguration

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.2 Benutzerprofile

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdspra-

chige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.2.1 Modell

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.2.2 Controller

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.2.3 Services

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.2.4 View

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.3 Volltextsuche

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.3.1 Modell

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.3.2 Controller

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest

gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.3.3 Services

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.3.4 View

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.4 Gewichtete Suche

Die gewichtete Suchanfragen werden mit *Boosting* realisiert. Wenn ein Suchbegriff zutrifft, wird die *Query*, die diesen Suchbegriff enthält, mit einem bestimmten Faktor geboosted. Damit ändert sich die Score und auch die Relevanz dieses Dokumentes gegenüber des anderen Dokumenten.

4.4.1 Modell

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.4.2 Controller

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.4.3 Services

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.4.4 View

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest

gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.5 Autovervollständigung

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.5.1 Modell

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.5.2 Controller

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdspra-

chige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.5.3 Services

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

4.5.4 View

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

5. Evaluierung

Während der Masterarbeit wurden mehrere *Search-Queries* implementiert, die aber aufgrund ihrer schlechten Effektivität in der Anwendung nicht angewendet werden. Da alle diese *Queries* durch schrittweise Optimierung von Nachteilen zu der Endergebnis geführt haben, werden auch diese *Queries* in diesem Kapitel evaluiert und detailliert beschrieben.

Um einzelne *Query* mit einander effektiv vergleichen zu können, müssen zwei Voraussetzungen erfüllt werden. Die Evaluierung muss genau auf die *Query* beschränkt werden und die Datenmenge muss unverändert bleiben. Der einzige Grund zu Änderung liegt in inkonsistenter Datenmenge, durch die die Evaluierungsergebnis stark beeinflusst wird. In der Tabelle unten wird deswegen der Datenausgangszustand kurz zusammengefasst.

Anzahl Dokumenten	Ältestes Datum	Neuestes Datum	Analyzer
3144	14.03.2015	11.10.2015	standard

Tabelle 5.1: Datenausgangszustand für die Evaluierung

Wie in dem vorherigen Kapitel schon erwähnt wurde, wird die Implementierung von allen *SearchQueries* in statische Methode separat ausgelagert und über eine Utility-Klasse (*SearchQueryUtility*) zur Verfügung gestellt. Für die Evaluierung werden diese Methode einzeln ausgeführt und die resultierte *SearchQuery* zusammen mit Pagination-Daten an *ElasticSearch*-Client direkt übergeben. Um die *SearchQuery* besonders die *Score*-Berechnung genau untersuchen zu können, bittet *ElasticSearch* durch *Explain*-Mode eine sehr detaillierte Ergebnisausgabe. Mit dem *SpringData-Framework* und *ElasticSearch-API* kann es wie folgt realisiert werden.

```
PageRequest pageRequest = new PageRequest(0, 50);
SearchQuery searchQuery = SearchQueryUtility.buildFullTextSearchBoolQuery_1(
    "Werkstudent Java, Datenbanken", pageRequest);
SearchResponse searchResponse = elasticsearchConfiguration.getNode()
    .client().prepareSearch("intellijob")
        .setQuery(searchQuery.getQuery())
        .setExplain(true)
        .setSize(50)
        .get();
```

5.1 Volltext-Suchanfragen

Query Nr. 1

```
{
  "bool": {
    "should": [
      {
        "match": {
          "content": {
            "query": "Werkstudent Java",
            "type": "boolean",
            "operator": "AND"
          }
        }
      },
      {
        "match": {
          "content": {
            "query": "Datenbanken",
            "type": "boolean"
          }
        }
      }
    ],
    "minimum_should_match": "1"
  }
}
```

Beschreibung: Es gibt zwei boolesche Suchanfragen, die zu einer gemeinsamen ebenfalls booleschen Suchanfrage zusammengefasst sind. Die Suchbegriffe sind mit *AND* verknüpft. „*minimum_should_match : 1*“ sorgt dafür, dass entweder die erste oder die zweite Suchanfrage treffen muss, damit das Dokument relevant wird.

#	Score	Arbeitsbezeichnung	Datum
1	0.13931298	WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK	21.09.2015 11-24
2	0.13591066	Werkstudent im Bereich Software Entwicklung - Smart HomeBosch Software Innovations GmbH	28.04.2015 02-50
3	0.13591066	Werkstudent im Bereich Software Entwicklung - Smart HomeBosch Software Innovations GmbH	28.04.2015 02-50

Tabelle 5.2: Query Nr. 1 - Ergebnis-Ausschnitt

Ergebnis: Zur Evaluierung vom ersten *Query* werden nur drei erste Ergebnisse angezeigt, die schon deutlich darauf hinweisen, dass die Datenmenge nicht konsis-

tent ist. Es gibt also mehrere gleiche Arbeitsangebote, die auch gleich gewichtet werden.

Query Nr. 1.1

Nachdem die duplizierte Arbeitsangebote gelöscht wurden, sind nur **2292** Arbeitsangebote übrig geblieben.

Ergebnis: Die Tabelle unten zeigt die ersten fünf Arbeitsangeboten.

#	Score	Arbeitsbezeichnung	Datum
1	0.19563562	Werkstudent im Bereich Software Entwicklung - Smart HomeBosch Software Innovations GmbH - Berlin	16.08.2015 02-29
2	0.1758025	Werkstudent im Bereich Softwareentwicklung (Java)	23.08.2015 02-39
3	0.16529423	WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK	21.09.2015 11-24
4	0.15650849	Werkstudent im Bereich Frontend Design	19.09.2015 03-13
5	0.15650849	Werkstudent im Bereich Software Development Toolchain	18.04.2015 02-53

Tabelle 5.3: Query Nr. 1.1 - Ergebnis-Ausschnitt

Um Effektivität von Query berechnen zu können, muss die Anzahl von relevanten Dokumenten für diese Suchanfrage bekannt werden. Deswegen wird in der Tabelle(5.4) die Relevanz von einzelnen Suchbegriffen zusammengefasst.

Suchbegriff	relevante Dokumente
Werkstudent	125
Java	305
Werkstudent {AND} Java	8
Datenbanken	171
Werkstudent {AND} Java {AND} Datenbanken	0

Tabelle 5.4: Relevanz von Suchbegriffen für die Volltextsuche

Aus der Tabelle(5.4) wird es ersichtlich, dass es insgesamt 179 relevante Arbeitsangebote für die aktuelle Suchanfrage gibt. Wenn diese Zahl zusammen mit der Ergebnisanzahl in die Formel zur Berechnung von *Recall* und *Precision* aus dem Kapitel 2.4.1 eingesetzt wird, kommt als Ergebnis raus:

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision
179	179	179	1 = 100%	1 = 100%

Tabelle 5.5: Query Nr. 1.1 - falsch-positive Precal, Precision

Auf den ersten Blick scheint die *Query* echt positiv zu sein. Denn die Relevanz eines Dokumentes bei der booleschen Suchanfrage wird allein durch das Existenz von gesuchten Termen bestimmt. Wenn man aber die Ergebnismenge aus der Tabelle(5.2) unter Berücksichtigung des Datums von zuletzt gespeicherten Arbeitsangebot am 11.10.2015 genau anschaut, fehlt es auf, dass einige Arbeitsangebote sehr veraltet sind und aus der Benutzersicht nicht mehr relevant sind. Für die Evaluierung wird deswegen die Relevanz von 10 best passenden Arbeitsangeboten subjektiv festgelegt.

#	Arbeitsbezeichnung	Datum	Rel.	Pos.	Ges.
1	Datenbankadministrator/-entwickler - Oracle	03.10.2015	10	10	100
2	IT-Consultant ERP/ Softwareentwickler	07.10.2015	9	9	81
3	WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK	21.09.2015	8	8	64
4	Junior Datenbank Analyst - Bereich Reise	02.10.2015	7	7	49
5	Werkstudent im Bereich Frontend Design	19.09.2015	6	6	36
6	Senior Java-Entwickler Angebotsimport	01.10.2015	5	5	25
7	Softwareentwickler Java für Versicherungen	30.09.2015	4	4	16
8	Java Entwickler DKB Service GmbH	24.09.2015	3	3	9
9	CONSULTANT BIG DATAT-Systems	21.09.2015	2	2	4
10	IT Consultant für Rollout-Projekte	01.10.2015	1	1	1

Tabelle 5.6: Die erwartende Ergebnismenge für Volltextsuche

Dabei wird nicht nur auf die gesuchte Terme sondern auch auf das Datum geachtet. Es werden sowohl die Dokumente als auch Ihre Positionierung in der Ergebnismenge von 10 sehr relevant bis 1 wenig relevant bewertet. Daraus ergibt sich, dass die Relevanz für den ersten Hit 100 Punkte ist.

$$Gesamt = Relevanz * Position \Rightarrow 10 * 10 = 100$$

Die maximale Relevanz für das Gesamtergebnis ergibt sich durch Summierung von alle 10 Hits und das ist genau 385 Punkte. Bei der Festlegung von Position innerhalb der erwartenden Ergebnisliste wurde auch auf die Anzahl von gesuchten Termen in Dokumenten geachtet. So beinhalten das Arbeitsangebot „Datenbankadministrator/-entwickler - Oracle“ 9. mal den gesuchten Term *Datenbank*. Außerdem ist das Arbeitsangebot ziemlich aktuell zu dem Datum 11.10.2015 und wird deswegen so hoch priorisiert. Die Dokumente „WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK“ und „Werkstudent im Bereich Frontend Design“ sind die einzelne erwartende Dokumente zu der Query *Werkstudent Java*. Da diese Query aus zwei mit AND verknüpften Termen besteht, werden diese Dokumente höher als Dokumente mit dem einzigen Term gewichtet. Aufgrund

ihres veralteten Datums werden sie jedoch nicht an der ersten und zweiten Stelle positioniert. Die restliche 6. Dokumente zu dem Suchterm *Werkstudent Java* sind noch älter und werden deswegen nicht in den ersten 10 erwartet. Unter Annahme von neuen Relevanz-Bedingungen werden die 10 besten Hits der ersten *Query* noch einmal berechnet.

#	Arbeitsbezeichnung	Datum	Ges.
1	Werkstudent im Bereich Software Entwicklung - Smart HomeBosch Software Innovations GmbH - Berlin	16.08.2015	0
2	Werkstudent im Bereich Softwareentwicklung (Java)	23.08.2015	0
3	WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK	21.09.2015	64
4	Werkstudent im Bereich Frontend Design	19.09.2015	36
5	Werkstudent im Bereich Software Development Toolchain	18.04.2015	0
6	Werkstudent/in Qualitätssicherung (automatisierte Softwaretests)	21.04.2015	0
7	Werkstudent/in Qualitätssicherung SeleniumMobility Media GmbH	20.08.2015	0
8	Werkstudent Personalberatung in Berlin-salesjob Personalberatung GmbH	01.05.2015	0
9	CONSULTANT BIG DATAT-Systems	21.09.2015	4
10	Java Entwickler DKB Service GmbH	24.09.2015	3

Tabelle 5.7: Query Nr. 1.1 - 10 beste Hits mit der Relevanzeinstufung

Das erste Hit in der Ergebnisliste gehört nicht zu den erwartenden Arbeitsangeboten und seine Relevanz ist deswegen 0. Seine Gesamt-Relevanz ist ebenfalls 0 ($Gesamt = Relevanz * Position \Rightarrow 0 * 10 = 0$). Das Arbeitsangebot „Java Entwickler DKB Service GmbH“ steht an der 10. Stelle, erwartet wird es aber an der 8. Stelle. Deswegen ist auch seine Gesamtrelevanz ($Gesamt = Relevanz * Position \Rightarrow 3 * 1 = 3$) kleiner geworden. Die finale Effektivität der ersten Query wird in der Tabelle(5.8) zusammengefasst.

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision	Ges.
10	4	10	40%	40%	107

Tabelle 5.8: Query Nr. 1.1 - finale Effektivitätsberechnung

Query Nr. 2

```
{
  "bool": {
    "should": [
      {
        "match": {
          "content": {
            "query": "Werkstudent Java",
            "type": "boolean",
            "operator": "AND"
          }
        }
      },
      {
        "match": {
          "content": {
            "query": "Datenbanken",
            "type": "boolean"
          }
        }
      }
    ],
    "minimum_should_match": "1"
  },
  "sort": [ { "receivedDate": { "order": "desc" } } ]
}
```

Beschreibung: Die zweite Query sortiert einfach die Ergebnisliste absteigend nach dem Feld *receivedDate*, das als Datum repräsentiert wird.

Ergebnis: Die Tabelle(5.9) zeigt, dass die Ergebnismenge jetzt absteigend nach Datum sortiert ist und alle Dokumente entweder den Suchterm *Werkstudent Java* oder *Datenbanken* beinhalten. *Recall* und *Precision* Werte sind zwar bei 40% geblieben, aber die Gesamt-Relevanz von 10 Hits ist um 9. Punkte hoch gestiegen. Daraus folgt, dass mindesten nach diesen Kriterien die zweite Query ein Stückchen besser als die erste Query ist.

Da die Ergebnisliste jetzt strikt nach Datum sortiert ist, tauchen zwei Arbeitsangebote „WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK“ und „Werkstudent im Bereich Frontend Design“ nicht auf. Trotz des drei wöchigen Unterschiedes können diese Arbeitsangebote immer noch aktuell sein. Außerdem sind es die einzige Arbeitsangebote, die zu dem Suchterm *Werkstudent Java* passen. Deswegen haben sie ihren Platz in 10 besten Hits durchaus verdient.

Interessant ist auch der *Explain*-Plan von dieser Query, dessen Ausschnitt im Anhang (siehe Anhang 5.1: Query Nr. 2 - Explain-Plan) zu finden ist. Die eigentliche Score-Berechnung für den ersten Hit ist **0.058416612** wird aber durch **NaN** ersetzt. **NaN** steht in *ElasticSearch* für maximale *Score*. Das heißt, dass die Dokumente, die eigentlich für die Suchanfrage nicht wirklich relevant sind, durch Sortierverfahren und das Maximale *Score* jedoch relevant werden. Daraus folgt,

#	Arbeitsbezeichnung	Datum	Ges.
1	IT-Consultant ERP/ Softwareentwickler	07.10.2015	90
2	Datenbankadministrator/-entwickler - Oracle	03.10.2015	90
3	Werkstudent / Studentische Aushilfe Human Resources KÖNIGSTEINER AGENTUR GmbH	03.10.2015	0
4	Rechtsanwalts- und Notarfachangestellte/r o.Ä. als Assistentin/Assistent für die Bereiche Revision und Recht/Beteiligungen	03.10.2015	0
5	Sekretär/in der Chefärztin Vivantes Humboldt-Klinikum - Berlin	03.10.2015	0
6	Junior Datenbank Analyst - Bereich Reise	02.10.2015	35
7	Qualitätsprüfer für den Warenein und -ausgang AneCom AeroTest GmbH	02.10.2015	0
8	Senior Softwareentwickler ASP.NET	01.10.2015	0
9	Systemingenieur Broadcast IT Systeme	01.10.2015	0
10	IT Consultant für Rollout-Projekte	01.10.2015	1

Tabelle 5.9: Query Nr. 2 - 10 beste Hits mit der Relevanzeinstufung

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision	Ges.
10	4	10	40%	40%	116

Tabelle 5.10: Query Nr. 2 - Effektivitätsberechnung

dass die *Score* bei der Sortierung nicht mehr berücksichtigt wird.

Query Nr. 3

```
{
  "function_score": {
    "query": {
      "bool": {
        "should": [
          {
            "match": {
              "content": {
                "query": "Werkstudent Java",
                "type": "boolean",
                "operator": "AND"
              }
            }
          }
        ],
        "minimum_should_match": 1
      }
    },
    "boost": 1
  }
}
```

```

        "content": {
          "query": "Datenbanken",
          "type": "boolean",
          "operator": "AND"
        }
      ]
    },
    "functions": [
      {
        "exp": {
          "receivedDate": {
            "scale": "14d"
          }
        }
      }
    ]
  }
}

```

Beschreibung: Die dritte *Query* ersetzt die Sortierung durch die *Decay*-Funktion (siehe Kapitel 2.5), um das Datum bei der *Score*-Berechnung zu berücksichtigen. Das Datum-Score wird jetzt *exponential* mit dem festgelegten *Scale*-Wert von 14 Tagen berechnet.

#	Arbeitsbezeichnung	Datum	Ges.
1	Junior Datenbank Analyst - Bereich Reise	02.10.2015	70
2	Sekretär/in der Chefärztin Vivantes Humboldt-Klinikum	03.10.2015	0
3	WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK	21.09.2015	64
4	Datenbankadministrator/-entwickler - Oracle	03.10.2015	70
5	Werkstudent im Bereich Frontend Design	19.09.2015	36
6	Senior Java-Entwickler Angebotsimport	01.10.2015	25
7	IT-Consultant ERP/ Softwareentwickler	07.10.2015	36
8	Werkstudent / Studentische Aushilfe Human Resources KÖNIGSTEINER	03.10.2015	0
9	Java Entwickler DKB Service GmbH - Berlin	24.09.2015	6
10	CONSULTANT BIG DATAT-Systems	21.09.2015	2

Tabelle 5.11: Query Nr. 3 - 10 beste Hits mit der Relevanzeinstufung

Ergebnis: Der *Explain*-Plan im Anhang (Anhang 5.2: Query Nr. 3 - Explain-Plan) zeigt schon deutlich, dass das Datum bei der *Score* mitgerechnet wird. Die neue Werte von *Recall*, *Recision* und Gesamt-Relevanz deuten auch darauf hin, dass die Ergebnismenge dank der *Decay*-Funktion viele besser geworden ist.

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision	Ges.
10	8	10	80%	80%	309

Tabelle 5.12: Query Nr. 3 - Effektivitätsberechnung

Query Nr. 4

```
{
  "function_score": {
    "query": {
      "bool": {
        "must": {
          "term": {
            "read": false
          }
        },
        "should": [
          {
            "match": {
              "content": {
                "query": "Werkstudent Java",
                "type": "boolean",
                "operator": "AND"
              }
            }
          },
          {
            "match": {
              "content": {
                "query": "Datenbanken",
                "type": "boolean",
                "operator": "AND"
              }
            }
          }
        ]
      }
    },
    "functions": [
      {
        "exp": {
          "receivedDate": {
            "scale": "14d",
            "decay": 0.5
          }
        }
      }
    ]
  }
}
```

Beschreibung: Die Query Nr. 4 erweitert die letzte Query mit einer neuen boole-
schen *Term-Query*, die dafür sorgt, dass nur ungelesene Arbeitsangebote durch-
gesucht werden.

Ergebnis: Interessanterweise hat diese winzige Änderung eine neue Ergebnisliste verursacht, was nicht erwartet und auch nie gewünscht wurde.

#	Arbeitsbezeichnung	Datum	Ges.
1	Sekretär/in der Chefärztin Vivantes Humboldt-Klinikum	03.10.2015	0
2	IT-Consultant ERP/ Softwareentwickler	07.10.2015	81
3	Junior Datenbank Analyst - Bereich Reise	02.10.2015	56
4	Datenbankadministrator/-entwickler - Oracle	03.10.2015	70
5	Senior Java-Entwickler Angebotsimport	01.10.2015	30
6	Qualitätsprüfer für den Warenein und -ausgang AneCom AeroTest	02.10.2015	0
7	Rechtsanwalts- und Notarfachangestellte/r o.Ä. als Assistentin/Assistent	03.10.2015	0
8	WERKSTUDENT SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK	21.09.2015	24
9	IT Consultant für Rollout-Projekte	01.10.2015	2
10	Softwareentwickler Java für Versicherungen	30.09.2015	4

Tabelle 5.13: Query Nr. 4 - 10 beste Hits mit der Relevanzeinstufung

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision	Ges.
10	8	10	70%	70%	267

Tabelle 5.14: Query Nr. 4 - Effektivitätsberechnung

Aus der Tabelle oben folgt, dass die neue Ergebnismenge im Vergleich zu dem alten Ergebnis schlechter geworden ist. Die Untersuchung von *Explain*-Plan (Anhang 5.3: Query Nr. 4 - Explain-Plan) hat gezeigt, dass die neue boolesche Term-*Query* bei der *Score*-Berechnung mitgezählt wird. Das darf aber nicht passieren. Das Feld *read* ist ein Flag, der ein Dokument aus dem Suchvorgang komplett ausschliessen soll und keinen Einfluss auf seine Relevanz haben darf.

Query Nr. 5

Beschreibung: Die fünfte und auch die letzte *Query* im Zusammenhang mit Volltextsuche verwendet ein Filter statt eine boolesche Term-Anfrage, um das Statusfeld *read* anzufragen. Das Filter kann jetzt die Dokumente nach dem *Read*-Status ohne Einfluss auf ihre Relevanz filtern.

Ergebnis: Die Ergebnismenge wird hier nicht extra aufgelistet, weil sie äquivalent zu dem Ergebnis der dritten *Query* ist. Der Explain-Plan von dem ersten Hit befindet sich im Anhang (Anhang 5.4: Query Nr. 5 - Explain-Plan).

```
{
  "function_score" : {
    "query" : {
      "filtered" : {
        "query" : {
          "bool" : {
            "should" : [ {
              "match" : {
                "content" : {
                  "query" : "Werkstudent Java",
                  "type" : "boolean",
                  "operator" : "AND"
                }
              }
            }, {
              "match" : {
                "content" : {
                  "query" : "Datenbanken",
                  "type" : "boolean",
                  "operator" : "AND"
                }
              }
            }
          ]
        }
      },
      "filter" : {
        "term" : {
          "read" : false
        }
      }
    },
    "functions" : [ {
      "exp" : {
        "receivedDate" : {
          "scale" : "14d"
        }
      }
    } ]
  }
}
```

5.2 Gewichtete Suchanfragen

Die gewichtete Suchanfrage verwendet als Suchterme die persönlichen Fähigkeiten und Qualifikationen, die der Benutzer selbst eingeschätzt hat. Diese Einschätzung muss jedenfalls bei der Suche berücksichtigt werden. Damit es sinnvoll getestet werden könnte, soll die Suchanfrage aus mehrere Suchbegriffen mit unterschiedlicher Einschätzung bestehen. Die *Query*-Struktur wird dadurch größer und wird deswegen in den Anhang ausgelagert.

Die Testszenarien beinhaltet vier unterschiedlich geschätzten Suchbegriffe. Die

Suchbegriffe aus der Volltextsuche sind schon bekannt und werden hier absichtlich nicht besonders hoch geschätzt.

Qualifikation	Einschätzung
Spring Framework	5
Java	3
PostgreSQL	2
Datenbanken	2

Tabelle 5.15: Einschätzung von Qualifikationen für die Evaluierung

Die Tabelle(5.16) zeigt auch die Anzahl von relevanten Dokumenten für die einzelnen Suchbegriffe.

Suchbegriff	relevante Dokumente
Spring	151
Framework	136
Spring {AND} Framework	43
Java	305
PostgreSQL	17
Datenbanken	171
Spring {AND} Framework {AND} Java {AND} PostgreSQL {AND} Datenbanken	2

Tabelle 5.16: Relevanz von Suchbegriffen für die gewichtete Suche

Die erwartende Ergebnismenge wird in der Tabelle(5.17) zusammengefasst.

#	Arbeitsbezeichnung	Datum	Rel.	Pos.	Ges.
1	Java Software Developer Connected Mobility	01.10.2015	10	10	100
2	Senior Java-Entwickler Angebotsimport	01.10.2015	9	9	81
3	Softwareentwickler (Java) für Banken	30.09.2015	8	8	64
4	Softwareentwickler Java für Versicherungen	30.09.2015	7	7	49
5	Java Developer! Inhouse Java Software Entwickler	28.09.2015	6	6	36
6	Software Developer (Java/Backend)	21.09.2015	5	5	25
7	Software Developer DWH	27.09.2015	4	4	16
8	Java Developer JEE / J2EE	03.10.2015	3	3	9
9	Junior Datenbank Analyst-Bereich Reise	02.10.2015	2	2	4
10	Softwareentwickler (Java) für Banken	30.09.2015	1	1	1

Tabelle 5.17: Die erwartende Ergebnismenge für die gewichtete Suche

Bei der Relevanzbestimmung wird auf das Vorkommen von Suchbegriffen in Abhängigkeit von ihrer Schätzung geachtet. Die maximale Relevanz für das Gesamt-

ergebnis beträgt ebenfalls 385 Punkte.

Query Nr. 6

Beschreibung: Die 6. Query (Anhang 5.5: Query Nr. 6) benutzt die *Boost*-Funktion, um die Score-Wert für einen bestimmten Suchterm mit dem *Boost factor* zu multiplizieren. Die *Boost factor* wird nach der Formel $Factor = Einschätzung * 10$ berechnet.

Ergebnis: Die Query findet nur fünf relevante Dokumente. Die Werte für *Recall* und *Recision* liegen also bei 50%. Da die relevanteste Dokumente jedoch gefunden werden, wird die Gesamtrelevanz der Ergebnismenge mit 291 Punkten trotzdem positiv bewertet. Aus dem *Explain*-Plan (Anhang 5.6: Query Nr. 6 - Explain-Plan von ersten Hit) wird es ersichtlich, dass nur ein exakter Suchtreffer geboosted wird. So werden nur die Begriffe, *Java* und *Datenbank*, aber nicht *Spring Framework* geboosted.

#	Arbeitsbezeichnung	Datum	Ges.
1	Senior Java-Entwickler Angebotsimport	01.10.2015	90
2	Softwareentwickler Java für Versicherungen	30.09.2015	63
3	Software Developer DWH	27.09.2015	32
4	Java Software Developer Connected Mobility	01.10.2015	70
5	Java Developer! Inhouse Java Software Entwickler	28.09.2015	36
6	Software Developer (PHP / Full Stack)	01.10.2015	0
7	Java Backend Entwickler	28.09.2015	0
8	Junior Softwareentwickler/in Java	07.09.2015	0
9	Senior Software Developer (Java)	21.09.2015	0
10	Java Frontend Developer	19.09.2015	0

Tabelle 5.18: Query Nr. 6 - 10 beste Hits mit der Relevanzeinstufung

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision	Ges.
10	5	10	50%	50%	291

Tabelle 5.19: Query Nr. 6 - Effektivitätsberechnung

Query Nr. 7

Beschreibung: Die letzte Query (Anhang 5.7: Query Nr. 7) verknüpft alle Suchterme mit *OR* und verwendet eine zusätzliche *Weight*-Funktion, um die Score noch einmal zu boosten. Im Zusammenhang mit *OR*-Verknüpfung werden jetzt alle Dokumente mit dem *Spring*- oder *Framework*-Term höher gewichtet.

Ergebnis: Obwohl die Ergebnisliste sich für 10 ersten Hits kaum geändert hat, ist die *Query* insgesamt besser geworden. So wird beispielsweise die Position des Arbeitsangebotes „*Softwareentwickler (Java) für Banken*“ um 10 Stellen (von 19 zu 9) geändert. Die *Query* hat allerdings auch einen Nachteil und zwar die Arbeitsangebote, wo das Wort *Framework* nicht im Zusammenhang mit *Spring* sondern mit *Play* oder *AngularJS* vorkommt, ebenfalls geboosted werden.

#	Arbeitsbezeichnung	Datum	Ges.
1	Senior Java-Entwickler Angebotsimport	01.10.2015	90
2	Softwareentwickler Java für Versicherungen	30.09.2015	63
3	Software Developer DWH	27.09.2015	32
4	Java Software Developer Connected Mobility	01.10.2015	70
5	Java Developer! Inhouse Java Software Entwickler	28.09.2015	36
6	Software Developer (PHP / Full Stack)	01.10.2015	0
7	Java Backend Entwickler	28.09.2015	0
8	Senior Software Developer (Java)	21.09.2015	0
9	Softwareentwickler (Java) für Banken	30.09.2015	2
10	Junior Softwareentwickler/in Java	07.09.2015	0

Tabelle 5.20: Query Nr. 7 - 10 beste Hits mit der Relevanzeinstufung

Ergebnis(<i>D</i>)	Rel. in Ergebnis (<i>a</i>)	Rel. in Dok(<i>S</i>)	Recall	Precision	Ges.
10	5	10	60%	60%	293

Tabelle 5.21: Query Nr. 7 - Effektivitätsberechnung

5.3 Vergleich und Fazit

Um Volltextsuche mit der gewichteten Suche sinnvoll vergleichen zu können, werden die *Query* Nr.5 und die *Query* Nr. 7 mit den gleichen Suchtermen(siehe Tabelle 5.15) ausgeführt. Die Werte für *Precall*, *Precision* und *Gesamt Relevanz* der Volltext-Suchanfrage sind äquivalent zu den Werten der 6. Query. Die Ergebnisliste sieht aber ein bisschen unterschiedlich aus(Tabelle 5.22). Die Untersuchung von Arbeitsangeboten „*Senior Software Developer (Java)*“ und „*Softwareentwickler (Java) für Banken*“ aus der gewichteten Suchanfrage hat ergeben, dass diese Arbeitsangebote aus der Benutzersicht viel relevanter als die Arbeitsangebote aus der Volltextsuche („*(Senior) PHP-Entwickler*“, „*Software Engineer Backend Zalando*“) sind. Daraus folgt, dass die gewichte Suche im Vergleich zur der Volltextsuche insgesamt ein besseres Ergebnis liefert.

#	Query	Arbeitsbezeichnung	Datum	Ges.
1	5	Senior Java-Entwickler Angebotsimport	01.10.2015	90
1	7	Senior Java-Entwickler Angebotsimport	01.10.2015	90
2	5	Softwareentwickler Java für Versicherungen	30.09.2015	63
2	7	Softwareentwickler Java für Versicherungen	30.09.2015	63
3	5	Software Developer DWH	27.09.2015	32
3	7	Software Developer DWH	27.09.2015	32
4	5	Java Software Developer Connected Mobility	01.10.2015	70
4	7	Java Software Developer Connected Mobility	01.10.2015	70
5	5	Java Developer! Inhouse Java Software Entwickler	28.09.2015	36
5	7	Java Developer! Inhouse Java Software Entwickler	28.09.2015	36
6	5	Software Developer (PHP / Full Stack)	01.10.2015	0
6	7	Software Developer (PHP / Full Stack)	01.10.2015	0
7	5	Java Backend Entwickler	28.09.2015	0
7	7	Java Backend Entwickler	28.09.2015	0
8	5	(Senior) PHP-Entwickler	21.09.2015	0
8	7	Senior Software Developer (Java)	21.09.2015	0
9	5	Software Engineer Backend Zalando	30.09.2015	0
9	7	Softwareentwickler (Java) für Banken	30.09.2015	2
10	5	Junior Softwareentwickler/in Java	07.09.2015	0
10	7	Junior Softwareentwickler/in Java	07.09.2015	0

Tabelle 5.22: Vergleich Query Nr. 5 und Query Nr. 7

6. Problemen und Schwierigkeiten

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

7. Zusammenfassung und Ausblick

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Literaturverzeichnis

- Dorschel, Joachim (2015). „Praxishandbuch Big Data - Wirtschaft – Recht – Technik“. 1. Aufl. Berlin Heidelberg New York: Springer-Verlag (siehe S. 4, 6).
- Jan, Steemann (2013). *Datenmodellierung in nicht relationalen Datenbanken*. url: <https://entwickler.de/online/datenbanken/datenmodellierung-in-nicht-relationalen-datenbanken-137872.html> (siehe S. 6).
- OnPage (2016). *Tag Cloud*. url: https://de.onpage.org/wiki/Tag_Cloud (siehe S. 12).
- Henrich, Andreas (2008). „Information Retrieval Grundlagen, Modelle und Anwendungen“. 1. Aufl. Otto-Friedrich-Universität Bamberg (siehe S. 12 f.).
- Mandl, Thomas (2001). „Tolerantes Information Retrieval“. Hochschulverband für Informatikwissenschaft(HI)e.V (siehe S. 14).
- Salton, G.; Wong, A.; Yang, C. S. (1975). „A Vector Space Model for Automatic Indexing“. *Commun. ACM* 18.11, S. 613–620. doi: 10.1145/361219.361220. url: <http://doi.acm.org/10.1145/361219.361220> (siehe S. 15).
- Rafal Kuc, Marek Rogozinski (2013). „Mastering Elasticsearch“. 1. Aufl. Packt Publishing Ltd. (siehe S. 20).
- Terrier, Francois (2013). *On Elasticsearch performance*. url: <https://blog.liip.ch/archive/2013/07/19/on-elasticsearch-performance.html> (siehe S. 20).

Abbildungsverzeichnis

2.1	Relationale Tabellen <i>skill-categories</i> und <i>skills</i>	4
2.2	Prozessablauf in IntelliJob	8
2.3	Schichtenarchitektur	10
2.4	Tag Clouds - <i>www.einfachbewusst.de</i>	11
2.5	Prozess des Wissenstransfers vom Autor eines Dokumentes bis zum Informations-Nachfragenden	13
2.6	Die elementaren booleschen Operationen	15
2.7	Vektorraumpräsentation eines Dokumentraums (Salton u. McGill 1987, 129)	16
2.8	Visualisierung von <i>Decay</i> -Funktion mit <i>Gaus</i>	19
2.9	Beispiel - Ein Cluster mit zwei Knoten, 4 <i>shards</i> und 1 <i>replica shard</i> [Terrier 2013]	20

Tabellenverzeichnis

2.1 Begriffsvergleich zwischen Relationale Datenbanken und Elasticsearch	18
5.1 Datenausgangszustand für die Evaluierung	30
5.2 Query Nr. 1 - Ergebnis-Ausschnitt	31
5.3 Query Nr. 1.1 - Ergebnis-Ausschnitt	32
5.4 Relevanz von Suchbegriffen für die Volltextsuche	32
5.5 Query Nr. 1.1 - falsch-positive Precal, Precision	32
5.6 Die erwartende Ergebnismenge für Volltextsuche	33
5.7 Query Nr. 1.1 - 10 beste Hits mit der Relevanzeinstufung	34
5.8 Query Nr. 1.1 - finale Effektivitätsberechnung	34
5.9 Query Nr. 2 - 10 beste Hits mit der Relevanzeinstufung	36
5.10Query Nr. 2 - Effektivitätsberechnung	36
5.11Query Nr. 3 - 10 beste Hits mit der Relevanzeinstufung	37
5.12Query Nr. 3 - Effektivitätsberechnung	38
5.13Query Nr. 4 - 10 beste Hits mit der Relevanzeinstufung	39
5.14Query Nr. 4 - Effektivitätsberechnung	39
5.15Einschätzung von Qualifikationen für die Evaluierung	41
5.16Relevanz von Suchbegriffen für die gewichtete Suche	41
5.17Die erwartende Ergebnismenge für die gewichtete Suche	41
5.18Query Nr. 6 - 10 beste Hits mit der Relevanzeinstufung	42
5.19Query Nr. 6 - Effektivitätsberechnung	42
5.20Query Nr. 7 - 10 beste Hits mit der Relevanzeinstufung	43
5.21Query Nr. 7 - Effektivitätsberechnung	43
5.22Vergleich Query Nr. 5 und Query Nr. 7	44

Glossar

Chunker - Machine Learning Technologie - Teilt Text in syntaktisch korrelierten Teile von Wörtern, wie Nomen Gruppen.

Coreference Resolution - Machine Learning Technologie - Bezugnahme auf dieselbe Entität.

DAO - Data Access Object - Adapter zur Abstrahierung und Entkopplung von Datenzugriffe.

Invertierter Index - wird in Suchserver eingesetzt und bezieht sich auf die Tatsache, dass nicht die Dokumente auf Worte zeigen, sondern eine Liste von Worten auf Dokumente zeigt.

IR - Information Retrieval

Named Entity Recognition - Machine Learning Technologie - erkennt und klassifiziert Bestandteile im Text.

NoSQL - not only SQL - bezeichnet Datenbanken, die einen nicht-relationalen Ansatz verfolgen.

Part-Of-Speech tagging - Machine Learning Technologie - Das Zuweisen von Markierungen zu einzelnen Einheiten (Wortart-Annotierung).

SentenceDetector - Machine Learning Technologie - teilt Text in einzelne Sätze

Tokenizer - Machine Learning Technologie - teilt Sätze in einzelne Worte

Anhang

Anhang 5.1: Query Nr. 2 - Explain-Plan

```
1 Hit: score(NaN) | receivedDate(07.10.2015 11-24) | name(IT-Consultant
      ERP/ Softwareentwickler (m/w)) | id(56727ceb44aeeb036a01c567)
```

```
1 Explain(: 0.058416612 = (MATCH) product of:
```

```
  0.116833225 = (MATCH) sum of:
```

```
    0.116833225 = (MATCH) weight(content:datenbank in 2248)
```

```
      [PerFieldSimilarity], result of:
```

```
        0.116833225 = score(doc=2248,freq=2.0), product of:
```

```
          0.5891623 = queryWeight, product of:
```

```
            3.5896857 = idf(docFreq=171, maxDocs=2292)
```

```
            0.16412643 = queryNorm
```

```
        0.19830398 = fieldWeight in 2248, product of:
```

```
          1.4142135 = tf(freq=2.0), with freq of:
```

```
            2.0 = termFreq=2.0
```

```
            3.5896857 = idf(docFreq=171, maxDocs=2292)
```

```
            0.0390625 = fieldNorm(doc=2248)
```

```
    0.5 = coord(1/2)
```

```
)
```

```
2 Hit: score(NaN) | receivedDate(03.10.2015 13-30) |
      name(Datenbankadministrator/-entwickler - Oracle (m/w)) |
      id(56727ce344aeeb036a01c565)
```

```
2 Explain(: 0.08178326 = (MATCH) product of:
```

```
  0.16356651 = (MATCH) sum of:
```

```
    0.16356651 = (MATCH) weight(content:datenbank in 2246)
```

```
      [PerFieldSimilarity], result of:
```

```
        0.16356651 = score(doc=2246,freq=2.0), product of:
```

```
          0.5891623 = queryWeight, product of:
```

```
            3.5896857 = idf(docFreq=171, maxDocs=2292)
```

```
            0.16412643 = queryNorm
```

```
        0.27762556 = fieldWeight in 2246, product of:
```

```
          1.4142135 = tf(freq=2.0), with freq of:
```

```
            2.0 = termFreq=2.0
```

```
            3.5896857 = idf(docFreq=171, maxDocs=2292)
```

```
            0.0546875 = fieldNorm(doc=2246)
```

```
    0.5 = coord(1/2)
```

```
)
```

Anhang 5.2: Query Nr. 3 - Explain-Plan

```

1 Hit: score(1.6383281E-4) | receivedDate(02.10.2015 03-18) | name(Junior
    Datenbank Analyst (m/w) - Bereich Reise) | id(56727b7544aeeb036a01c502)
1 Explain(: 1.6383281E-4 = (MATCH) function score, product of:
    0.10016363 = (MATCH) product of:
        0.20032726 = (MATCH) sum of:
            0.20032726 = (MATCH) weight(content:datenbank in 2147)
                [PerFieldSimilarity], result of:
                    0.20032726 = score(doc=2147,freq=3.0), product of:
                        0.5891623 = queryWeight, product of:
                            3.5896857 = idf(docFreq=171, maxDocs=2292)
                            0.16412643 = queryNorm
                        0.3400205 = fieldWeight in 2147, product of:
                            1.7320508 = tf(freq=3.0), with freq of:
                                3.0 = termFreq=3.0
                            3.5896857 = idf(docFreq=171, maxDocs=2292)
                            0.0546875 = fieldNorm(doc=2147)
                    0.5 = coord(1/2)
            0.0016356517 = (MATCH) Math.min of
                0.0016356517 = (MATCH) Function for field receivedDate:
                    0.0016356517 = exp(- MIN[Math.max(Math.abs(1.443748724E12(=doc value)
                        - 1.454944683321E12(=origin))) - 0.0(=offset), 0)] *
                        5.73038343716886E-10)
                3.4028235E38 = maxBoost
            1.0 = queryBoost
        )
-----
2 Hit: score(1.6062698E-4) | receivedDate(03.10.2015 03-16) |
    name(Sekretaer/in der ChefaerztinVivantes Humboldt-Klinikum - Berlin) |
    id(56727c3444aeeb036a01c552)
2 Explain(: 1.6062697E-4 = (MATCH) function score, product of:
    0.09346658 = (MATCH) product of:
        0.18693316 = (MATCH) sum of:
            0.18693316 = (MATCH) weight(content:datenbank in 2227)
                [PerFieldSimilarity], result of:
                    0.18693316 = score(doc=2227,freq=2.0), product of:
                        0.5891623 = queryWeight, product of:
                            3.5896857 = idf(docFreq=171, maxDocs=2292)
                            0.16412643 = queryNorm
                        0.31728637 = fieldWeight in 2227, product of:
                            1.4142135 = tf(freq=2.0), with freq of:
                                2.0 = termFreq=2.0
                            3.5896857 = idf(docFreq=171, maxDocs=2292)
                            0.0625 = fieldNorm(doc=2227)
                    0.5 = coord(1/2)
            0.0017185498 = (MATCH) Math.min of
                0.0017185498 = (MATCH) Function for field receivedDate:

```

TABELLENVERZEICHNIS

```
0.0017185498 = exp(- MIN[Math.max(Math.abs(1.443835E12(=doc value) -
1.454944683321E12(=origin))) - 0.0(=offset), 0)] *
5.73038343716886E-10)
3.4028235E38 = maxBoost
1.0 = queryBoost
)
-----
3 Hit: score(1.5946852E-4) | receivedDate(21.09.2015 11-24) |
name(WERKSTUDENT (M/W) SOFTWAREENTWICKLUNG IM BEREICH BAHNTECHNIK) |
id(56727c7e44aeeb036a01c564)
3 Explain(: 1.594685E-4 = (MATCH) function score, product of:
0.16529422 = (MATCH) product of:
0.33058843 = (MATCH) sum of:
0.33058843 = (MATCH) sum of:
0.2607186 = (MATCH) weight(content:werkstudent in 2245)
[PerFieldSimilarity], result of:
0.2607186 = score(doc=2245,freq=5.0), product of:
0.638943 = queryWeight, product of:
3.892993 = idf(docFreq=126, maxDocs=2292)
0.16412643 = queryNorm
0.40804672 = fieldWeight in 2245, product of:
2.236068 = tf(freq=5.0), with freq of:
5.0 = termFreq=5.0
3.892993 = idf(docFreq=126, maxDocs=2292)
0.046875 = fieldNorm(doc=2245)
0.06986981 = (MATCH) weight(content:java in 2245)
[PerFieldSimilarity], result of:
0.06986981 = score(doc=2245,freq=1.0), product of:
0.49461055 = queryWeight, product of:
3.0135949 = idf(docFreq=305, maxDocs=2292)
0.16412643 = queryNorm
0.14126226 = fieldWeight in 2245, product of:
1.0 = tf(freq=1.0), with freq of:
1.0 = termFreq=1.0
3.0135949 = idf(docFreq=305, maxDocs=2292)
0.046875 = fieldNorm(doc=2245)
0.5 = coord(1/2)
9.647555E-4 = (MATCH) Math.min of
9.647555E-4 = (MATCH) Function for field receivedDate:
9.647555E-4 = exp(- MIN[Math.max(Math.abs(1.442827456E12(=doc value) -
1.454944683321E12(=origin))) - 0.0(=offset), 0)] *
5.73038343716886E-10)
3.4028235E38 = maxBoost
1.0 = queryBoost
)
-----
4 Hit: score(1.435483E-4) | receivedDate(03.10.2015 13-30) |
name(Datenbankadministrator/-entwickler - Oracle (m/w)) |
id(56727ce344aeeb036a01c565)
```

TABELLENVERZEICHNIS

```
4 Explain(: 1.4354828E-4 = (MATCH) function score, product of:
  0.08178326 = (MATCH) product of:
    0.16356651 = (MATCH) sum of:
      0.16356651 = (MATCH) weight(content:datenbank in 2246)
        [PerFieldSimilarity], result of:
          0.16356651 = score(doc=2246,freq=2.0), product of:
            0.5891623 = queryWeight, product of:
              3.5896857 = idf(docFreq=171, maxDocs=2292)
              0.16412643 = queryNorm
            0.27762556 = fieldWeight in 2246, product of:
              1.4142135 = tf(freq=2.0), with freq of:
                2.0 = termFreq=2.0
              3.5896857 = idf(docFreq=171, maxDocs=2292)
              0.0546875 = fieldNorm(doc=2246)
          0.5 = coord(1/2)
      0.0017552283 = (MATCH) Math.min of
        0.0017552283 = (MATCH) Function for field receivedDate:
          0.0017552283 = exp(- MIN[Math.max(Math.abs(1.443871853E12(=doc value)
            - 1.454944683321E12(=origin))) - 0.0(=offset), 0]) *
            5.73038343716886E-10)
        3.4028235E38 = maxBoost
      1.0 = queryBoost
    )
```

```
-----
5 Hit: score(1.3447095E-4) | receivedDate(19.09.2015 03-13) |
  name(Werkstudent (m/w) im Bereich Frontend Design) |
  id(5672639144aeeb036a01bdbe)
5 Explain(: 1.3447093E-4 = (MATCH) function score, product of:
  0.15650848 = (MATCH) product of:
    0.31301695 = (MATCH) sum of:
      0.31301695 = (MATCH) sum of:
        0.21985723 = (MATCH) weight(content:werkstudent in 287)
          [PerFieldSimilarity], result of:
            0.21985723 = score(doc=287,freq=2.0), product of:
              0.638943 = queryWeight, product of:
                3.892993 = idf(docFreq=126, maxDocs=2292)
                0.16412643 = queryNorm
              0.3440952 = fieldWeight in 287, product of:
                1.4142135 = tf(freq=2.0), with freq of:
                  2.0 = termFreq=2.0
                3.892993 = idf(docFreq=126, maxDocs=2292)
                0.0625 = fieldNorm(doc=287)
            0.093159735 = (MATCH) weight(content:java in 287)
              [PerFieldSimilarity], result of:
                0.093159735 = score(doc=287,freq=1.0), product of:
                  0.49461055 = queryWeight, product of:
                    3.0135949 = idf(docFreq=305, maxDocs=2292)
                    0.16412643 = queryNorm
                  0.18834968 = fieldWeight in 287, product of:
```

TABELLENVERZEICHNIS

```
1.0 = tf(freq=1.0), with freq of:
  1.0 = termFreq=1.0
  3.0135949 = idf(docFreq=305, maxDocs=2292)
  0.0625 = fieldNorm(doc=287)
0.5 = coord(1/2)
8.5919263E-4 = (MATCH) Math.min of
8.5919263E-4 = (MATCH) Function for field receivedDate:
  8.5919263E-4 = exp(- MIN[Math.max(Math.abs(1.442625233E12(=doc value)
    - 1.454944683321E12(=origin))) - 0.0(=offset), 0]) *
    5.73038343716886E-10)
3.4028235E38 = maxBoost
1.0 = queryBoost
)
```

Anhang 5.3: Query Nr. 4 - Explain-Plan

```

1 Hit: score(3.9707313E-4) | receivedDate(03.10.2015 03-16) |
      name(Sekretaeer/in der ChefaerztinVivantes Humboldt-Klinikum - Berlin) |
      id(56727c3444aeeb036a01c552)
1 Explain(: 3.9707313E-4 = (MATCH) function score, product of:
  0.23496056 = (MATCH) product of:
    0.35244083 = (MATCH) sum of:
      0.16806811 = (MATCH) weight(read:F in 2227) [PerFieldSimilarity],
        result of:
          0.16806811 = score(doc=2227,freq=1.0), product of:
            0.16494422 = queryWeight, product of:
              1.0189391 = idf(docFreq=2248, maxDocs=2292)
              0.16187838 = queryNorm
            1.0189391 = fieldWeight in 2227, product of:
              1.0 = tf(freq=1.0), with freq of:
                1.0 = termFreq=1.0
              1.0189391 = idf(docFreq=2248, maxDocs=2292)
              1.0 = fieldNorm(doc=2227)
            0.18437272 = (MATCH) weight(content:datenbank in 2227)
              [PerFieldSimilarity], result of:
                0.18437272 = score(doc=2227,freq=2.0), product of:
                  0.5810925 = queryWeight, product of:
                    3.5896857 = idf(docFreq=171, maxDocs=2292)
                    0.16187838 = queryNorm
                  0.31728637 = fieldWeight in 2227, product of:
                    1.4142135 = tf(freq=2.0), with freq of:
                      2.0 = termFreq=2.0
                    3.5896857 = idf(docFreq=171, maxDocs=2292)
                    0.0625 = fieldNorm(doc=2227)
                0.6666667 = coord(2/3)
            0.0016899565 = (MATCH) Math.min of
              0.0016899565 = (MATCH) Function for field receivedDate:
                0.0016899565 = exp(- MIN[Math.max(Math.abs(1.443835E12(=doc value) -
                  1.454973962236E12(=origin))) - 0.0(=offset), 0)] *
                  5.73038343716886E-10)
              3.4028235E38 = maxBoost
            1.0 = queryBoost
  )

```

Anhang 5.4: Query Nr. 5 - Explain-Plan

```

1 Hit: score(1.6048997E-4) | receivedDate(02.10.2015 03-18) | name(Junior
    Datenbank Analyst (m/w) - Bereich Reise) | id(56727b7544aeeb036a01c502)
1 Explain(: 1.6048997E-4 = (MATCH) function score, product of:
    0.10016363 = (MATCH) product of:
        0.20032726 = (MATCH) sum of:
            0.20032726 = (MATCH) weight(content:datenbank in 2147)
                [PerFieldSimilarity], result of:
                    0.20032726 = score(doc=2147,freq=3.0), product of:
                        0.5891623 = queryWeight, product of:
                            3.5896857 = idf(docFreq=171, maxDocs=2292)
                            0.16412643 = queryNorm
                        0.3400205 = fieldWeight in 2147, product of:
                            1.7320508 = tf(freq=3.0), with freq of:
                                3.0 = termFreq=3.0
                                3.5896857 = idf(docFreq=171, maxDocs=2292)
                                0.0546875 = fieldNorm(doc=2147)
                    0.5 = coord(1/2)
            0.0016022779 = (MATCH) Math.min of
                0.0016022779 = (MATCH) Function for field receivedDate:
                    0.0016022779 = exp(- MIN[Math.max(Math.abs(1.443748724E12(=doc value)
                        - 1.454980658242E12(=origin))) - 0.0(=offset), 0]) *
                        5.73038343716886E-10)
                3.4028235E38 = maxBoost
            1.0 = queryBoost
    )

```

Anhang 5.5: Query Nr. 6

```
{
  "function_score": {
    "query": {
      "filtered": {
        "query": {
          "bool": {
            "should": [
              {
                "match": {
                  "content": {
                    "query": "Spring Framework",
                    "type": "boolean",
                    "operator": "AND",
                    "boost": 50.0
                  }
                }
              },
              {
                "match": {
                  "content": {
                    "query": "Java",
                    "type": "boolean",
                    "operator": "AND",
                    "boost": 30.0
                  }
                }
              },
              {
                "match": {
                  "content": {
                    "query": "PostgreSQL",
                    "type": "boolean",
                    "operator": "AND",
                    "boost": 20.0
                  }
                }
              },
              {
                "match": {
                  "content": {
                    "query": "Datenbanken",
                    "type": "boolean",
                    "operator": "AND",
                    "boost": 20.0
                  }
                }
              }
            ]
          }
        }
      }
    }
  },
  "filter": {
    "term": {
      "read": false
    }
  }
}
```

```

    }}}
  },
  "functions": [
    {
      "exp": {
        "receivedDate": {
          "scale": "14d"
        }
      }
    }
  ]
}

```

Anhang 5.6: Query Nr. 6 - Explain-Plan von ersten Hit

```

1 Hit: score(5.763883E-4) | receivedDate(01.10.2015 02:55) | name(Senior
    Java-Entwickler Angebotsimport (m/w)Idealo Internet GmbH - Berlin) |
    id(567264c744aeeb036a01be31)
1 Explain(: 5.763883E-4 = (MATCH) function score, product of:
0.39150882 = (MATCH) product of:
0.52201176 = (MATCH) sum of:
0.39096564 = (MATCH) sum of:
0.17041172 = (MATCH) weight(content:spring in 403)
    [PerFieldSimilarity], result of:
0.17041172 = score(doc=403,freq=2.0), product of:
0.5933848 = queryWeight, product of:
    3.7132995 = idf(docFreq=151, maxDocs=2292)
    0.15979987 = queryNorm
0.28718585 = fieldWeight in 403, product of:
    1.4142135 = tf(freq=2.0), with freq of:
        2.0 = termFreq=2.0
    3.7132995 = idf(docFreq=151, maxDocs=2292)
    0.0546875 = fieldNorm(doc=403)
0.22055392 = (MATCH) weight(content:framework in 403)
    [PerFieldSimilarity], result of:
0.22055392 = score(doc=403,freq=3.0), product of:
0.609988 = queryWeight, product of:
    3.8171992 = idf(docFreq=136, maxDocs=2292)
    0.15979987 = queryNorm
0.36157092 = fieldWeight in 403, product of:
    1.7320508 = tf(freq=3.0), with freq of:
        3.0 = termFreq=3.0
    3.8171992 = idf(docFreq=136, maxDocs=2292)
    0.0546875 = fieldNorm(doc=403)
0.06734424 = (MATCH) weight(content:java^30.0 in 403)
    [PerFieldSimilarity], result of:
0.06734424 = score(doc=403,freq=2.0), product of:
0.28894326 = queryWeight, product of:
    30.0 = boost
    3.0135949 = idf(docFreq=305, maxDocs=2292)
    0.0031959976 = queryNorm

```

TABELLENVERZEICHNIS

```
0.23307082 = fieldWeight in 403, product of:
  1.4142135 = tf(freq=2.0), with freq of:
    2.0 = termFreq=2.0
    3.0135949 = idf(docFreq=305, maxDocs=2292)
    0.0546875 = fieldNorm(doc=403)
0.06370189 = (MATCH) weight(content:datenbank^20.0 in 403)
  [PerFieldSimilarity], result of:
0.06370189 = score(doc=403,freq=2.0), product of:
  0.22945254 = queryWeight, product of:
    20.0 = boost
    3.5896857 = idf(docFreq=171, maxDocs=2292)
    0.0031959976 = queryNorm
  0.27762556 = fieldWeight in 403, product of:
    1.4142135 = tf(freq=2.0), with freq of:
      2.0 = termFreq=2.0
      3.5896857 = idf(docFreq=171, maxDocs=2292)
      0.0546875 = fieldNorm(doc=403)
  0.75 = coord(3/4)
0.0014722231 = (MATCH) Math.min of
  0.0014722231 = (MATCH) Function for field receivedDate:
    0.0014722231 = exp(- MIN[Math.max(Math.abs(1.443660906E12(=doc value)
      - 1.455040566297E12(=origin))) - 0.0(=offset), 0]) *
      5.73038343716886E-10)
    3.4028235E38 = maxBoost
  1.0 = queryBoost
)
```

Anhang 5.7: Query Nr. 7

```
{
  "function_score": {
    "query": {
      "filtered": {
        "query": {
          "bool": {
            "should": [
              {
                "function_score": {
                  "query": {
                    "match": {
                      "content": {
                        "query": "Spring Framework",
                        "type": "boolean",
                        "operator": "OR",
                        "boost": 50.0
                      }
                    }
                  }
                },
                "functions": [
                  {
                    "weight": 50.0
                  }
                ]
              }
            ]
          }
        },
        "functions": [
          {
            "weight": 30.0
          }
        ]
      }
    },
    "functions": [
      {
        "weight": 30.0
      }
    ]
  },
  {
    "function_score": {
      "query": {
        "match": {
          "content": {
            "query": "Java",
            "type": "boolean",
            "operator": "OR",
            "boost": 30.0
          }
        }
      },
      "functions": [
        {
          "weight": 30.0
        }
      ]
    },
    {
      "function_score": {
        "query": {
          "match": {
            "content": {
              "query": "PostgreSQL",
```

```

        "type": "boolean",
        "operator": "OR",
        "boost": 20.0
      }}
    },
    "functions": [
      {
        "weight": 20.0
      }
    ]
  },
  {
    "function_score": {
      "query": {
        "match": {
          "content": {
            "query": "Datenbanken",
            "type": "boolean",
            "operator": "OR",
            "boost": 20.0
          }
        }
      },
      "functions": [
        {
          "weight": 20.0
        }
      ]
    }
  }
]
},
"filter": {
  "term": {
    "read": false
  }
}
},
"functions": [
  {
    "exp": {
      "receivedDate": {
        "scale": "14d"
      }
    }
  }
]
}

```

Anhang 5.8: Query Nr. 7 - Explain-Plan von ersten Hit

```

1 Hit: score(0.024797508) | receivedDate(01.10.2015 02-55) | name(Senior
      Java-Entwickler Angebotsimport (m/w)Idealo Internet GmbH - Berlin) |
      id(567264c744aeeb036a01be31)

```

TABELLENVERZEICHNIS

1 Explain(: 0.024797508 = (MATCH) function score, product of:
17.131985 = (MATCH) product of:
22.842648 = (MATCH) sum of:
19.548283 = (MATCH) function score, product of:
0.39096564 = (MATCH) sum of:
0.17041172 = (MATCH) weight(content:spring in 403)
[PerFieldSimilarity], result of:
0.17041172 = score(doc=403,freq=2.0), product of:
0.5933848 = queryWeight, product of:
3.7132995 = idf(docFreq=151, maxDocs=2292)
0.15979987 = queryNorm
0.28718585 = fieldWeight in 403, product of:
1.4142135 = tf(freq=2.0), with freq of:
2.0 = termFreq=2.0
3.7132995 = idf(docFreq=151, maxDocs=2292)
0.0546875 = fieldNorm(doc=403)
0.22055392 = (MATCH) weight(content:framework in 403)
[PerFieldSimilarity], result of:
0.22055392 = score(doc=403,freq=3.0), product of:
0.609988 = queryWeight, product of:
3.8171992 = idf(docFreq=136, maxDocs=2292)
0.15979987 = queryNorm
0.36157092 = fieldWeight in 403, product of:
1.7320508 = tf(freq=3.0), with freq of:
3.0 = termFreq=3.0
3.8171992 = idf(docFreq=136, maxDocs=2292)
0.0546875 = fieldNorm(doc=403)
50.0 = (MATCH) Math.min of
50.0 = (MATCH) product of:
1.0 = constant score 1.0 - no function provided
50.0 = weight
3.4028235E38 = maxBoost
1.0 = queryBoost
2.0203273 = (MATCH) function score, product of:
0.06734424 = (MATCH) weight(content:java^30.0 in 403)
[PerFieldSimilarity], result of:
0.06734424 = score(doc=403,freq=2.0), product of:
0.28894326 = queryWeight, product of:
30.0 = boost
3.0135949 = idf(docFreq=305, maxDocs=2292)
0.0031959976 = queryNorm
0.23307082 = fieldWeight in 403, product of:
1.4142135 = tf(freq=2.0), with freq of:
2.0 = termFreq=2.0
3.0135949 = idf(docFreq=305, maxDocs=2292)
0.0546875 = fieldNorm(doc=403)
30.0 = (MATCH) Math.min of
30.0 = (MATCH) product of:
1.0 = constant score 1.0 - no function provided

TABELLENVERZEICHNIS

```
    30.0 = weight
    3.4028235E38 = maxBoost
    1.0 = queryBoost
1.2740378 = (MATCH) function score, product of:
    0.06370189 = (MATCH) weight(content:datenbank^20.0 in 403)
        [PerFieldSimilarity], result of:
    0.06370189 = score(doc=403,freq=2.0), product of:
    0.22945254 = queryWeight, product of:
        20.0 = boost
        3.5896857 = idf(docFreq=171, maxDocs=2292)
        0.0031959976 = queryNorm
    0.27762556 = fieldWeight in 403, product of:
        1.4142135 = tf(freq=2.0), with freq of:
            2.0 = termFreq=2.0
            3.5896857 = idf(docFreq=171, maxDocs=2292)
            0.0546875 = fieldNorm(doc=403)
20.0 = (MATCH) Math.min of
    20.0 = (MATCH) product of:
        1.0 = constant score 1.0 - no function provided
        20.0 = weight
        3.4028235E38 = maxBoost
        1.0 = queryBoost
0.75 = coord(3/4)
0.0014474393 = (MATCH) Math.min of
    0.0014474393 = (MATCH) Function for field receivedDate:
        0.0014474393 = exp(- MIN[Math.max(Math.abs(1.443660906E12(=doc value)
            - 1.455070193609E12(=origin))) - 0.0(=offset), 0)] *
            5.73038343716886E-10)
        3.4028235E38 = maxBoost
        1.0 = queryBoost
)
```

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt.

Datum: 10. Februar 2016, Berlin

Unterschrift: